English Library:
the Linguistics Bookshelf

Volume 1

Aurelia Martelli and Virginia Pulcini
(eds)

Investigating English with Corpora

Studies in Honour of Maria Teresa Prat

Polimetrica
International Scientific Publisher

Open Access Publications

The publication of this book was made possible with funding from the Dipartimento di Scienze del Linguaggio e Letterature Moderne e Comparate, Università degli Studi di Torino.

2008 Polimetrica ® S.a.s.
Corso Milano, 26
20052 Monza – Milano – Italy
Phone ++39. 039.2301829
Web site: www.polimetrica.com

ISSN 1974-0603 Printed Edition
ISSN 1974-0670 Electronic Edition
ISBN 978-88-7699-103-5 Printed Edition
ISBN 978-88-7699-104-2 Electronic Edition

Table of Contents

1. Corpora and Discourse Analysis

2. Corpora in Lexicology and Lexicography

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

3. Corpora and Translation

4. Corpora and Specialized Discourse

5. Corpora and English Language Teaching

6. Corpora and Historical Studies

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

Tabula Gratulatoria

Donatella Abbate Badin
Laurie Anderson
Simona Anselmi
Guy Aston
Pablo Luis Avila
Julia Bamford
Ljiljana Banjanin
Manuel Barbera
Paul Bayley
Carla Bazzanella
Gaetano Berruto
Antonio Bertacca
Paolo Bertinetti
Anna Bianco
Cecilia Boggio
Rosa Maria Bollettieri Bosinelli
Marina Bondi
Derek Boothman
Sandra Bosco
Anna Brawer
Nicholas Brownlees
Paola Brusasco
Silvia Bruti
Annamaria Caimi
M. Cristina Caimotto
Antonella Calogiuri
Bona Cambiaghi
Sandra Campagna
Umberto Capra
Pinuccia Caracchi
Melita Cataldi
Toni Cerutti
Anna Chiarloni
Stephen Coffey
Michelangelo Conoscenti
Mikaela Cordisco
Giuseppina Cortese
Marcella Costa
Pietro Deandrea
Caterina D’Elia
Gabriella Del Lungo Camiciotti
Tullio De Mauro
Giancarlo De Pretis
Bruna Di Sabato
Vittoria Dolcetti Corazza
Gerard Dorrity
Marina Dossena
John Douthwaite
Rosanna Ducati
Richard Dury
Roberta Facchinetti
Giuliana Ferreccio
Silvana Ferreri
Valerio Fissore
Lucia Folena
Cristiano Furiassi
Mario e Bice Garavelli
Laura Gavioli
Sara Gesuato
Fedora Giordano
Maurizio Gotti
Sylviane Granger
Ruth Anne Henderson
Shan Hirst
Giovanni Iamartino
Susan Kermas

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

Barbara Lanati
Liliana Landolfi
Sara Laviosa
Clarinda Lawry
Lucilla Lopriore
Fiona MacWilliam
Brunello Mantelli
Carla Marello
Flavia e Franco Marenco
Mariagrazia Margarito
Aurelia Martelli
Lorenzo Massobrio
Gerardo Mazzaferro
Davide Mazzi
Gabriella Mazzon
Lavinia Merlini
Donna R. Miller
Maria Isabella Mininni
Vincenza Minutella
Alessandra Molino
Filippo Monge
John Morley
Mariangela Mosca
Amanda Murphy
Andrea Nava
Stefania Nuccorini
Elana Ochse
Renato Oliva
Christopher Owen
Luisa Pantaleoni

Nella Panzarasa
Alan Partington
Maria Pavesi
Luciana Pedrazzini
Giorgio Pestelli
Marco Piovaz
Vanda Polese
Donatella Ponti
Graziella Pozzo
Virginia Pulcini
Davide Ricca
Maria Cecilia Rizzardi
G. Matteo Roccati
Rita Salvi
Elisabetta Soletti
Martin Solly
Joanna Spendel
Mario Squartini
Tullio Telmon
Elena Tognini-Bonelli
Paola Tornaghi
Claudia Tresso
Flo Ulmann
Margherita Ulrych
Carla Vaglio Marengo
Paolo e Nicoletta Valabrega
Nicoletta Vasta
Jocelyne Vincent
Barbara Zandrino
Federico Zanettin

Introduction

Aurelia Martelli and Virginia Pulcini

The first and foremost reason for compiling this volume is to honour the long and brilliant career of our colleague Maria Teresa Prat, Full Professor of *Lingua Inglese* at the *Facoltà di Lingue e Letterature Straniere* of the University of Turin. *Investigating English with Corpora* has been chosen as the title and the unifying theme of this *Festschrift* to focus on a field of research – corpus linguistics – in which she has been involved, as testified by the volume on learner corpora published in 2004 and by more recent articles about the use of corpora as a resource for teachers and students. The contributors to the volume are friends and colleagues from the University of Turin and from other Italian universities who kindly accepted to write about aspects of their ongoing research related to the principles and methodology of corpus linguistics. Thus, this volume also offers an interesting snapshot of research areas and topics being investigated with the support of corpus data at the time of writing.

In the course of her intense activity as a scholar and lecturer in English language and linguistics, which is still 'in full swing', Maria Teresa Prat's contribution has been remarkable, both in terms of scientific production and in her active presence in academic and institutional settings at national and international level. The titles of her publications included in this book show a variety of research interests ranging from the teaching and learning of English, linguistic education and curriculum design at school and university levels, English grammar, lexis, and corpus linguistics, spanning over a period of more than thirty years.

In some of her writings she deliberately chose to describe her career experience as marked by the significant changes which affected the academic scene from the late 1960s to the present time. In *The Study of English Language in Italian Universities* (1991), for example, she pictures the gradual transformation of the role of

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

English in the Italian university starting from "What we were…" and moving on to "What we are…" and to "What the future holds for us…". The use of the inclusive 'we' meaningfully refers to herself and the first generation of academics who witnessed and actively participated in these transformations in the Italian university system and in its degree curricula. A personal recollection of the radical turn-around in the study of English from the late 1960s onwards is taken up again in *Computer Learner Corpora* (2004).

The first major change which she recollects back in 1967, when she graduated in English Literature at the University of Turin, is the transformation of the university from an élite place of learning, primarily focussed on the appreciation of literature, to a higher education institution which masses of students from different backgrounds flooded into, expecting to be trained in foreign languages both for educational and professional purposes. Hence the need to cater for a wide audience of Italian learners of foreign languages, the urgent demand for a new vision of learning goals and methodologies, the development of materials and the provision of adequate resources, such as language centres, and – most centrally – the appointment of expert language staff. We cannot but remember the serious problems posed by the recruitment and the institutional role of the *lettori* or "foreign language experts" in the Italian university system – a longstanding debate to which Maria Teresa Prat contributed with great energy, balance and sense of duty.

In her view of foreign language teaching, she has always given great importance to mutual cooperation with teachers and trainers working in the secondary school context, as well as to sharing problems and experiences with colleagues of foreign languages other than English, in adherence to the principles of multilingualism and multiculturalism. What stands out significantly in her attitude to our job as linguists in higher education is the beneficial support of linguistic research to the improvement of related areas such as foreign language pedagogy. Her model academic profile reflects both the components of coordinator and observer of practical language instruction – a huge amount of work which often "goes unnoticed and unrecorded" (Prat Zagrebelsky 1991: 7) – and lecturer and scholar of English language and linguistics.

The second major event which we wish to recall in this brief overview of Maria Teresa Prat's career – which reflects an academic path shared by many scholars of her generation – is the achieved independence of *Lingua Inglese* as an academic discipline in its own right, against a tradition in which language teaching courses were mainly confined to a subordinate or ancillary role within the field of literary studies. The University of Turin was among the first in Italy to implement a new model based on the introduction of *Lingua Inglese* as a compulsory subject with independent status with respect to *Letteratura Inglese* in the traditional curriculum of Modern Languages. At least in the local dimension of the degree course in Modern Languages in the *Facoltà di Lettere e Filosofia*, where Maria Teresa Prat became Associate Professor of *Lingua Inglese* in 1982, and from 1997 in the *Facoltà di Lingue e Letterature Straniere*, where she became Full Professor in 2004, this model has always combined a balanced component of practical language competence and a theoretical and descriptive approach to the study of English language and translation. In the wake of major cultural and institutional events which took place in the 1980s and 1990s in the field of European education – the growing importance of foreign languages, opportunities for student mobility, the 'Bologna Process' and the '3+2' reform of the university system in Italy introduced in 1998 – the number of posts of *Lingua Inglese* in Italian universities has progressively and substantially grown from a handful of scholars in the 1980s to a host of researchers and professors in 'language and literature' and non-linguistic faculties, even though a numerical balance between 'language' and 'literature' professors has not yet been achieved.

Inspired by a strong interest in innovation in both its institutional and educational dimensions, Maria Teresa Prat's research has focussed on some major areas of English linguistics and English language teaching. Her most important books deal with the notion of grammar and its role in linguistic education (*Grammatica e lingua straniera*, 1985), English dictionaries and their use for language learning (*Dal dizionario ai dizionari*, 1989; *Guida all'uso del dizionario inglese-italiano*, 1997), English language teaching (*L'inglese per capire. Guida alla lettura e alla traduzione di testi* *delle scienze umane*, 1997) and English vocabulary (*Lessico e apprendimento linguistico*, 1998). Moreover, a strong involvement in English teaching, teacher training (*Guida all'aggiornamento degli insegnanti*, 1987), linguistic education and curriculum design in schools and university (*The Study of English Language in Italian Universities*, 1991) and the importance of language centres emerge very clearly from her constant and generous contribution of articles to journals, conference proceedings and books throughout her career.

In the 1990s she became interested in corpus linguistics and was involved in the ICLE (International Corpus of Learner English) as coordinator of the Italian component. This international project, coordinated by the Centre for English Corpus Linguistics (CECL), *Université catholique de Louvain*, consisted in the collection of learner English produced by students from different language backgrounds. This led to the publication of *Computer Learner Corpora* (2004) focussed on the Italian component of the ICLE. In the introduction to this volume Maria Teresa Prat describes her involvement in corpus linguistics as a 'pioneering' exploration into a completely new discipline. In her interesting personal recollection, which she calls "a Pilgrim's Progress" from literature to corpus linguistics, she refers again to her own training and academic journey from a tradition rooted in historical, philological and literary approaches to English studies to a gradual appreciation of 20th century 'mainstream' linguistics, and later on to the discovery of the electronic medium and the strong potential of computer corpora as a methodology to describe and teach the English language in its manifold aspects.

The 'corpus revolution' – a major turning point in linguistics – has introduced a new approach to the study of English and the manipulation and interpretation of linguistic data. A substantial, first-class contribution has been given by Italian scholars, working in the areas of both corpus linguistics and computational linguistics.[1] As for English, a body of scholars from within the *Associazione Italiana di Anglistica* (A.I.A.), who are well-represented in this volume, have

[1] Among the contributions given by the scholars of the University of Turin in corpora and linguistic engineering, see the volume M. Barbera, E. Corino and C. Onesti (eds) (2007), *Corpora e linguistica in rete*, Guerra, Perugia.

produced advanced research of international reputation in many fields of linguistic enquiry, thus establishing a tradition of which Italian academia can be proud. We are pleased to celebrate Maria Teresa Prat as one of the founding ‘mothers’ of this tradition.

***

The studies included in this volume are focussed on ‘corpus data’ of different types, from collections of language samples, smaller *ad hoc* specialized corpora, to already existing large corpora such as the British National Corpus. They are grouped around six major research areas, namely Discourse Analysis, Lexicology and Lexicography, Translation, Specialized Discourse, English Language Teaching and Historical Studies. Within each section they are arranged following the authors’ names in alphabetical order. A short abstract of each article is presented in the following thematic sections.

1. Corpora and Discourse Analysis

In “Weakness and fear: a fragment of corpus-assisted discourse analysis”, **Paul Bayley** (University of Bologna) shows how the methods of corpus linguistics, when applied to small specialized corpora, can be profitably combined with discourse analysis. The corpus used is a collection of political speeches made by US Presidents between 1960 and 2004. Starting from the assumption that symbolic oppositions are a central feature of US political rhetoric, the analysis focuses on the two antonymous pairs *WEAKNESS* and *STRENGTH* and *FEAR* and *COURAGE*. Their frequency, instances of co-selection and semantic patterns in the corpus are initially discussed. Then the analysis concentrates on the semantic value of *WEAKNESS* and *FEAR* occurring together in a stretch of text which is a passage from J.F. Kennedy’s inaugural address in 1960. The contextual, co-texual and intertextual elements of discourse are examined, and the negative value associated with these two words within the context of the Cold War and the nuclear arms race is clearly highlighted.

The discourse of children’s rights with regard to the so-called “street children” is addressed by **Giuseppina Cortese** (University of Turin) in "Personal narratives in children's rights discourse". Working from a constructivist notion of children as capable members of local subcultures, whose basic rights appear to be dramatically violated, this study aims to give them voice. A corpus of web-distributed accounts is examined both manually and through concordancing, with a focus on the polarizations in children's relationships to adults emerging from the professional literature on the implementation of the *Convention on the Rights of the Child*. Based on the Labovian approach to narrative syntax as well as on more recent narratological perspectives within a broad pragmalinguistic view of language as action, findings confirm that violence and abuse are perpetrated on these particularly vulnerable children within the family context and at the hands of the police. Particularly with reference to the master narrative of the nurturing mother, children's narratives configure a counter-narrative.

In "The illocutionary force of interrogatives in English varieties", **Roberta Facchinetti** (University of Verona) highlights what types of illocutionary acts are preferably carried out by means of interrogatives in English and if and to what extent such illocutionary acts can be generalized among different varieties of English or rather (partly) differ from one variety to another. The analysis is focussed on the four mental state verbs – *KNOW*, *SEE*, *THINK*, and *WANT* – which are among the most frequent lexical verbs in English. Their occurrences in collocation with central modal verbs (e.g. "may I know…?" and "can't you see…?") in the seven components of the International Corpus of English currently available are examined. In interrogative contexts, the modalized mental state verbs under scrutiny appear to be relatively limited in number, with some Asian varieties exhibiting the highest frequencies. A semantic-pragmatic analysis of the data highlights the following recurrent illocutionary values: request for information, request for action, and rhetorical question. Such uses are discussed in detail with reference both to their distribution and to their syntactic realizations.

In "The perception of citizenship in the English press", **John Morley** (University of Siena) and **Alan Partington** (University of Bologna) analyse a part of the Pilot Corpus of texts from two English newspapers, namely the *Guardian* (left-wing, broadsheet) and the *Sun* (right-wing, tabloid), collected from the week of

Monday 21st to Saturday 26th November 2006, as part of the EU-sponsored IntUne project on the perception of citizenship in Europe in the English press. The authors discover that it is of little use looking ‘directly’ for abstract keywords such as *PERCEPTION* or *CITIZENSHIP*, given their rarity in newspaper language and the tendency of the Anglo-American press to personalize and reify even political issues. However, they are able to throw considerable light ‘indirectly’ on perceptions of belonging and identity by examining and comparing items such as *EUROPE*/*EUROPEAN*, as well as by compiling frequency lists from the newspapers and examining them for what they call “topic words”, which reveal some of the particular political, social and economic preoccupations of each newspaper. One important finding is that, although the popular and the quality press report the ‘same’ news stories, their styles of doing so remain different in spite of frequently made allegations about the so-called ‘tabloidization’ of the quality press.

The discourse of job advertising is investigated by **Martin Solly** (University of Florence) in “Job ads and the construction of identity in contemporary English primary education”. He analyses a small corpus of job advertisements placed by English state sector primary schools in the same 2007 issue of the *Times Educational Supplement* – for many years the most important employment forum in English education – in order to study the ways in which the schools define and shape their own identity. The author illustrates a range of techniques that the print ads use to attract and recruit potential candidates, yet at the same time to define the individual school’s identity and project it to the outside world. The dichotomy between the shared features in the recruitment ads and the schools’ need to compete with each other in order to attract the most suitable applicants gives rise to considerable differences within what might be expected to be a comparatively limited (sub-) genre. Furthermore, since many of the schools use the ads to state their current and future policies and ambitions, they also reveal much about the current nature of primary school education in England.

2. Corpora in Lexicology and Lexicography

In "Remarks on the frequency and phraseology of *A/AN* in Modern English", **Stephen Coffey** (University of Pisa) carries out a corpus-based investigation on the indefinite article, addressing the question of how frequently it is used as an independent function word or, by contrast, as part of lexico-phraseological units and frames. The data, extracted from the British National Corpus, consist of three separate samples of 500 random tokens of *A*, together with three 500-token samples taken from the spoken sub-corpus and one 500-token sample of *AN*. The author proposes a categorization of uses of the indefinite article *A/AN* including numerical phrases (*A MILE AWAY*), quantifiers (*A LOT*), phrases modifying a noun (*AS A WHOLE*, *QUITE A LOT*) and several types of functional phrases and expressions. The results show that over 80% of occurrences of *A/AN* in the written samples from the BNC are independent function words, while a smaller proportion includes instances of phrasal units. In the spoken samples the proportion of *A/AN* as an independent function word amounts to about 60%, to a higher proportion of phrasal units, as well as repetitions and incompletions typical of the spoken language.

Valerio Fissore and **Ruth Anne Henderson** (University of Turin) investigate the structural complexity of the English noun phrase, which is both an asset of the language and a hurdle for the non-native speaker, in "Disambiguation of English pre- and postmodified noun phrases". The first part introduces the use of noun phrase premodification in English and some problems the learner of English encounters when resorting to this elegant and economical syntactic strategy of the English language. The second section looks at both pre- and postmodification of the noun phrase, with reference to scientific and medical texts and to the problem of disambiguation. The authors show that, when a new notion is produced, the head of the noun phrase is likely to be massively modified on its right-hand side with notional or material relations by means of expressed prepositions, relative pronouns or other connectors. Once the notion has established itself, the information originally contained in postmodification is moved to the left-hand side of the head, while the linkers are dropped. This structure of the

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

noun phrase is often baffling for the translator into a romance language. Disambiguation may only occur if the process can be reversed to reconstruct the transformational path, and move from pre- to postmodification.

In "What dictionaries leave out: new non-adapted Anglicisms in Italian", **Cristiano Furiassi** (University of Turin) describes procedures followed to verify whether some non-adapted Anglicisms, which are not included in recent editions of Italian monolingual dictionaries, may be retrieved in a corpus of Italian newspaper language including texts written before the year 2000, i.e. the La Repubblica Corpus. The procedures combine n-gram based identification of English words in the Italian language, automatic extraction of wordlists containing non-adapted Anglicisms from the corpus, and semi-automatic refinement of the wordlists obtained with the aid of a spell-checker, i.e. OpenOffice Writer. The manually refined list of non-adapted Anglicisms extracted from the corpus is cross-checked against an exclusion corpus, i.e. De Mauro and Mancini (2003). Finally, a selection of non-adapted Anglicisms, which are absent from recent editions of Italian monolingual dictionaries but which the author believes should reasonably be included, is presented. Although the overall approach is mainly quantitative, some non-adapted Anglicisms are also qualitatively analysed.

In "'Phraseologies' and Italian-English dictionaries: evidence for a proposal", **Stefania Nuccorini** (Roma Tre University) analyses a few examples of supposed true friends, pairs of cognate words in English and in Italian which, despite their overlapping etymology and meaning, are often used in different syntagmatic patterns in the two languages. The pairs discussed are *ASSOLUTAMENTE/ABSOLUTELY*, *REALE/REAL*, *TERRORISTA/TERRORIST* and *KAMIKAZE/KAMIKAZE*. Often their respective collocational behaviours, semantic prosodies and phraseologies do not match their lexico-semantic friendship and as a consequence they cannot be considered as translational equivalents. The analysis starts from the (mis-)use of these supposed true friends in the Italian component of the International Corpus of Learner English (ICLE) and then concentrates on their lexicographic treatment in two bilingual dictionaries and in a monolingual English dictionary. The author argues that the case studies reported offer enough evidence for the systematic analysis of the phraseological behaviour of other pairs. She also suggests that relevant information about mismatching phraseologies should be included in bilingual dictionaries to the benefit, in particular, of advanced learners engaged in production activities.

The linguistic information extracted from corpora in the course of the ongoing compilation of a *Dizionario di anglicismi* is described by **Virginia Pulcini** (University of Turin) in "Corpora and lexicography: the case of a dictionary of Anglicisms". Two Italian corpora – the La Repubblica Corpus and the itWaC – representing newspaper and web language respectively – have been searched in order to check the frequency, orthographic form, meaning, uses and aspects of the lexical profile of the Italian Anglicisms to be included in the dictionary. The corpora have also been exploited to verify whether Anglicisms are more or less frequent than their Italian equivalents, if they exist. The British National Corpus is used to check whether the selected Anglicisms correspond to English words (or are in fact false Anglicisms) and to what extent their lexical behaviour is the same. The article claims that, on the one hand, corpus evidence is indispensable for modern dictionary-making but, on the other hand, corpora, far from being all-inclusive, have various limitations imposed by what they contain (topics, registers) and the period of time that they cover. Therefore it is essential for the lexicographer to balance the data with his/her own native speaker intuition and with the criteria designed for the specific type of dictionary and its prospective users.

3. Corpora and Translation

In "Interjections in translated Italian: looking for traces of dubbed language", **Silvia Bruti** (University of Pisa) and **Maria Pavesi** (University of Pavia) deal with the translation of interjections from English into Italian in a corpus of dubbed films. The authors compare the frequency and functions of interjections in the source language and in the target language in order to better situate the third code, i.e. the translated film language. They then provide an analysis of primary interjections and show that interjections in dubbing exhibit an atypical distribution in comparison to spoken

Italian. Their analysis shows that interjections in dubbed Italian are globally less frequent and less varied than in spontaneous speech and that the frequency patterns from the film translation corpus differ from the corresponding patterns from the spoken Italian corpus, thus suggesting the self-standing status of dubbed Italian. Moreover, the authors point to the strong influence of English on this ‘variety’ of translated Italian as shown by the fact that Italian interjections which exhibit a degree of similarity with English ones tend to be over-represented in dubbing, whereas interjections which are specific and restricted to Italian tend to be under-represented.

In “Corpus studies and translation universals: a critical appraisal”, **Sara Laviosa** (University of Bari) discusses the notion of universals of translation, an issue which has been widely debated from a theoretical and methodological viewpoint, particularly since the advent of corpus-based translation studies. The analysis highlights the contribution of current investigations to the progress of the discipline as a whole. The author starts from Baker’s original definition of translation universals as linguistic features which typically occur in translated rather than original texts and are thought to be independent of the influence of the specific language pairs. The perspective adopted is product and target-oriented since Baker proposes investigating translation universals – simplification, explicitation, normalisation or conservatism, levelling out and distinctive distribution of target-language items – by comparing translated texts *vis-à-vis* comparable texts written originally in the target language. Other translation scholars within the descriptive school, most notably Toury and Chesterman, have put forward divergently similar definitions and typologies of universals. The author appraises the uniformity and diversity of corpus studies of translation universals in terms of divergent similarity, as intended by Chesterman, namely the sum of relevant samenesses and differences among entities originating from the same unity.

In “‘What’s in a name?’ References to women in *Romeo and Juliet* and their translation into Italian”, **Vincenza Minutella** (University of Turin) examines the words used to refer to women in Shakespeare’s play, using opensourceshakespeare.org, an online searchable corpus of Shakeaspeare’s complete works. She focusses in particular on the word *MAID* in its singular, plural and combined forms, as this lexical item appears to be the most morphologically productive and the most amenable to ambiguity and to wordplay in *Romeo and Juliet*. The word MAID is analysed in its co-text and context of use, and its combinatory patterns and meanings, especially with negative connotation as regards sex and violence. Then the author considers the Italian equivalents for the word MAID chosen by different translators of this tragedy for the page and for the stage. She discusses some problematic issues involved in translating a theatre text and points out some repeated or common patterns in the translation of words referring to women, which reflect broader translation tendencies. The author explains that the reasons for these tendencies are to be found in the target context in which the translations are produced, and in the medium of transmission and the function of the target text in the receiving culture.

In “Towards a corpus-based distinction between language-specific and universal features of mediated discourse”, **Margherita Ulrych** (Catholic University of Milan) and **Simona Anselmi** (Catholic University of Piacenza) argue that mediated discourse encompasses a wide variety of linguistic forms ranging from the translation, editing or revision of texts by native speakers to the use of a lingua franca or an interlanguage by non-native speakers. Corpus-based studies of these various types of mediated discourse indicate they share some common features irrespective of the native languages involved. Studies in translation, ELF and learner English have all investigated the possible presence of universal features over and above the specific languages in contact. A crucial element in verifying these findings, however, is to differentiate universal from language-specific factors. To this end, this paper examines instances of ELF and translational features in a parallel corpus of texts written in English by non-native speakers within the EU and then edited by native speakers with a background in translation. The results are then compared to errors found in learner corpora.

4. Corpora and Specialized Discourse

In “Telling a convincing story: a corpus-assisted analysis of business presentations”, **Julia Bamford** (University of Naples, l’Orientale) examines a small specialized corpus of oral business presentations to find which discursive practices are used by the speakers to present a positive and convincing picture of the company's performance. The author examines some of the most frequently occurring rhetorical strategies, in particular those that illustrate the personalization and conversationalization of business discourse and the use of narratives in presentations. In order to investigate these aspects the author focusses on the use of first person pronouns and identifies some patterns in the use of *I* and *WE* which emerge from the observation of concordances. Another aspect investigated is the presentation of both good and bad news, a crucial component of this type of business presentation, often to be found embedded in good news in order to minimize the impact. The methodology used, which involves concordancing and ethnographic interviews, as well as the specialized nature of the corpus, allow the author to carry out a fine-grained analysis of the presentations and to examine critically the techniques used to convince and persuade and show how using corpora to analyse specialized discourse can help identify its characterizing features.

In "'In this article, we focus on…': metadiscourse across disciplines", **Marina Bondi** and **Davide Mazzi** (University of Modena and Reggio Emilia) deal with the notion of metadiscourse which has become central in studies on academic discourse and its genres, both from a cross-disciplinary and a cross-linguistic perspective. The paper explores the frequency and use of key metadiscourse items in English research article openings in the fields of economics and history. Data show that metadiscourse has two main functional roles in both disciplines. First, it introduces the purpose of the paper and/or it represents the broader research field, thus situating the study within the relevant disciplinary debate and setting up a dialogue with the discourse community. Secondly, it may either signal the author's voice or reported argumentation, thus laying emphasis on the interplay of argumentative voices as a constitutive factor in the construction of disciplinary discourse. As for the overall epistemological configurations revealed by collocational and phraseological patterns, the findings show that in the field of economics openings appear more centred on the discovery, observation and experimentation of empirical facts and the evaluation of their impact and underlying factors. In history, by contrast, objects of the discipline, such as events, trends and discontinuities, tend to be more theoretically discussed, conceptualised and set against the appropriate paradigmatic background.

The article by **Sandra Campagna** (University of Turin), “Refreshing the globe? A corpus-based study of ‘corporate ethos’”, is an analysis of promotional homepages representative of global companies operating in the food and beverage sector. The author uses corpus-assisted techniques to expand on previous studies investigating how global/local identities are semiotically constructed in corporate culture through the interface of the Web. The data analysed consist of two small language corpora respectively crawled from the Coca-Cola Company website and from the McDonald’s Company website. The comparison with a larger reference corpus allows the author to identify a range of keywords contextualizing ‘glocalization’ and ‘ethos’. A selection of these terms is then analysed through concordancing in order to find how the keywords defining ‘glocalization’ and ‘ethos’ behave in context. The author then compares her own corpus-assisted findings with claims advanced in previous studies on the semiotic construction of ‘glocalization’ in corporate culture. Data in both corpora used by the author confirm the claims of ‘glocalization’ advanced in the previous studies on the multimodal representation of corporate culture and display a strong correlation between corporate culture and ethical responsibility.

In “CADIS – A Corpus of Academic Discourse”, **Maurizio Gotti** (University of Bergamo) focuses on a corpus of texts for academic communication (CADIS), specially compiled by CERLIS, the research centre on specialized languages based at the University of Bergamo. The corpus comprises texts for academic communication written in English, produced by scholars and academic institutions in various parts of the world. It also comprises some Italian texts for comparative purposes. This paper concentrates on the criteria followed in the construction of such a specialised digital corpus and discusses aspects pertaining to the collection and classification of texts according to their genre, the gathering of ethnographic data on the communicative events and actors involved, and a reconstruction of the general and specific sociolinguistic context of such texts. All these factors are expected to contribute to the definition of identity variants and their evaluation and interpretation are being carried out in the light of the latest research insights.

5. Corpora and English Language Teaching

This section opens with the essay “It’s only human…”, in which **Guy Aston** (University of Bologna) looks at some implications of a more ‘forgiving’ approach to learner corpus analysis. The author argues that the analysis of learner corpora has traditionally been based on comparison with native-speaker norms with the identification and enumeration of features which appear underused, overused, or simply wrong. This has implied a notion of error which evaluates the quality of particular features in polar terms (right-wrong) rather than in degrees of pragmatic effectiveness (better-worse). In this paper the author questions the focus on native speaker norms in learner corpus analysis, as it ignores a range of factors which – from a teaching and learning perspective – should be considered of equal if not greater importance. Drawing on corpus data from the British National Corpus, Aston reminds us that to err is not only human but also strategic and that in the approach to learner corpora, the adoption of a more forgiving approach may contribute to developing means of analysis which cast greater light on the phenomenon of the successful foreign language learning.

Umberto Capra (University of Piemonte Orientale) discusses how corpus linguistics has brought an impressive change to our understanding of language in use in “Keeping the corpus-based promise to language teaching in schools: the need for a quantum leap”. The combination of new data processing tools with the revolutionary global hypertext of the World Wide Web has offered an unprecedented opportunity to collect and analyse corpora. There have been high expectations about the contribution that corpus linguistics can make to language teaching and learning. Yet, while corpus-based approaches strongly influenced language education and training at university level, language teaching in schools seems to have profited much less from such seminal points of view. Reasons for this lack of influence range from teacher education to the traditional role played by (open choice) grammar in the language description to learners’ attitudes, motivations and learning strategies. The author argues that it may be a cognitive question, similar to the one facing physics teachers when they try to get their students, groping through a traditionally mechanical world, to grasp quantum physics.

In “Error analysis and learner corpora: a study of errors in the written production by English students of Italian”, **Aurelia Martelli** (University of Turin) explores the possibility of combining a corpus-based approach with the principles of error analysis in order to study the performance of learners of a foreign language. The data used is a 30,000-word corpus of learner Italian collected at Lancaster University during the 1998-1999 academic year and tagged for errors. Errors are initially analysed in terms of their frequency to allow the identification of error categories whose frequency emerges as statistically significant. Subsequently such categories are analysed more closely by looking at individual errors in the linguistic context in which they occur. In spite of the controversies that characterise error annotation, it is claimed that this procedure represents a new valuable approach to the study of interlanguage and that error analysis can greatly benefit from the use of the tools and methods of a computer-based approach to language.

In “The Role-Play Learner Corpus: a resource for investigating learner language”, **Maria Cecilia Rizzardi**, **Luciana Pedrazzini** and **Andrea Nava** (University of Milan) explore the use of learner corpora as a resource to investigate learner language and to improve foreign language teaching and testing techniques. The first part of the paper provides a short update on recent developments in the learner corpus research field and reviews the main methods of eliciting and collecting data. The authors then describe a project carried out at the University of Milan to compile a small-scale corpus of spoken learner language (the Role-Play Learner Corpus) and discuss some issues related to the analysis of this Corpus. The final section includes some considerations on the role of learner corpus research in foreign language pedagogy and possible implications for teacher development.

6. Corpora and Historical Studies

In “19CSC, second-generation corpora and the history of English”, **Marina Dossena** and **Richard Dury** (University of Bergamo) outline the methodological choices made by the authors in the preparation of a Corpus of Nineteenth-century Scottish Correspondence (19CSC), currently being compiled at the University of Bergamo. The authors explain how the corpus is expected to meet the requirements of ‘second-generation’ corpora, so that it may be complementary to other projects being carried out elsewhere in Europe. After a short introduction on the concept of ‘second-generation’ corpora, the structure and contents of 19CSC are discussed, and preliminary findings are outlined. For instance, the authors point out how different discourse communities and social networks interact quite dynamically in the documents contained in the corpus and change and adopt different communicative strategies as social or geographical distance increases, or mutual acquaintance becomes more firmly established. The concluding section discusses pathways for future research.

In “Language change and variation in English: the case of unsplit *FOR TO* in infinitival purpose clauses”, **Gerardo Mazzaferro** (University of Turin) offers a corpus-based historical investigation of unsplit *FOR TO* used to introduce infinitival purpose clauses. Drawing on both diachronic and synchronic corpora of English, the author considers the origins and development of *FOR TO* from the early Middle English period, when it was used to mark purpose infinitives, to the early Modern English period and shows that after the 17th century its use has become linguistically, socio-culturally and geographically restricted. The study points out that a limited number of occurrences can also be found in EFL and ESL varieties of present-day English such as Indian English or Philippine English. The author concludes that there is no agreement about the origins of *FOR TO* and points to two main positions: the first affirms that some English dialects have inherited *FOR TO* from western Middle English; while the second affirms that the development of *FOR TO* is probably due to external factors, namely the Scandinavian and French or Anglo-Norman presence in England.

***

We would like to warmly thank all the colleagues who have agreed to contribute to this book. The wide range of research areas covered and the projects presented in this collection of essays testify the great interest in corpus linguistics of Italian scholars and their valuable contribution to this field of research at national and international levels.

We are also grateful to Giovanni Iamartino for accepting this volume in the new series *English Library: the Linguistics Bookshelf* of Polimetrica. We warmly thank Giuseppina Cortese for her guidance and advice in the preparation of this volume. We are also indebted to Elana Ochse, Fiona MacWilliam and Christopher Owen for their help in linguistic revisions and Cristiano Furiassi and Vincenza Minutella for their help in the final proofreading.

Finally, we wish to express our deepest gratitude to Maria Teresa Prat – Micia – , for her academic and human qualities and for setting for us a constant example of sense of duty, professionalism, loyalty, strength, optimism and generosity in the course of our respective careers and in years to come.

January 2008

Publications by Maria Teresa Prat

Books:

(1985) *Grammatica e lingua straniera*, La Nuova Italia, Firenze.

(1997) *Guida all'uso del dizionario inglese-italiano*, Zanichelli, Bologna.

(1997) *L'inglese per capire. Guida alla lettura e alla traduzione di testi delle scienze umane*, UTET Libreria, Torino.

Edited volumes:

(1974) (ed.) *Introduzione e note a Short Stories di W. Somerset Maugham*, Società Editrice Internazionale, Torino.

(1986) [1982] (ed.) *Avviamento alla lettura in lingua inglese. Dispense per gli studenti del 1° anno*, Giappichelli Editore, Torino. (Introduction by M.T. Prat Zagrebelsky, pp. 3-8).

(1987) (ed. with M.L. Caccia, L. Fontanella, G. Mortarotto and M. Negarville) *Guida all'aggiornamento degli insegnanti*, *CIRDA*, Università di Torino.

(1989) (ed.) *Dal dizionario ai dizionari. Orientamento e guida all'uso per studenti di lingua inglese*, Tirrenia Stampatori, Torino.

(1991) (ed.) *The Study of English Language in Italian Universities*, Edizioni dell'Orso, Alessandria, (Introduction by M.T. Prat Zagrebelsky, pp. xiii-xx).

(1998) (ed.) *Lessico e apprendimento linguistico. Nuove tendenze della ricerca e pratiche didattiche*, La Nuova Italia, Firenze (Introduction by M.T. Prat Zagrebelsky, pp. ix-xi).

(2004) (ed.) *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage*, Edizioni dell'Orso, Alessandria.

Articles in journals, books, conference proceedings:

(1971) "Botta Vincenzo", in A.M. Ghisalberti, M. Pavan, F. Bartoccini (eds), *Dizionario biografico degli italiani,* Vol. 13, Istituto della Enciclopedia Italiana, Roma, pp. 379-380.

(1977) "Reclutamento e aggiornamento del personale insegnante", *Cooperazione Educativa* 3, pp. 107-110.

(1978) "Alcune considerazioni sul problema dell'insegnamento della letteratura italiana in riferimento alla riforma della scuola secondaria", in M.G. Caponera and C. Siani (eds), *La letteratura di lingua straniera nella secondaria superiore,* Zanichelli, Bologna, pp. 15-20.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

(1980) “Il posto della lettura nella scuola media. Scelta e preparazione dei materiali”, in G. Cortese (ed.), *La lettura nelle lingue straniere. Aspetti teorici e pratici,* Franco Angeli, Milano, pp. 459-483.

(1981) “Un’attività comunicativa di lettura, facilmente realizzabile in classe: Utilizzo della corrispondenza in inglese nella scuola media”, *LEND. Lingua e Nuova Didattica* X (2), pp. 33-37.

(1982) “Il ruolo dell’università nella formazione in servizio degli insegnanti, con particolare riferimento al settore dell’anglistica”, in E. Siciliani, R. Barone and G. Aston (eds), *La lingua inglese nell’università. Linee di ricerca, esperienze, proposte,* pp. 243-255.

(1982) “Parlato e scritto”, in M.T. Prat Zagrebelsky (ed.), *Avviamento alla lettura in lingua inglese. Dispense per gli studenti del 1° anno*, Giappichelli Editore, Torino, pp. 9-31.

(1983) “Common core e grammatica pedagogica ai livelli iniziali dell’apprendimento dell’inglese all’università”, in M.P. De Angelis, V. Fortunati and V. Poggi (eds), *Atti del V convegno dell’Associazione Italiana di Anglistica,* CLUEB, Bologna, pp. 375-382.

(1983) “Criteri per la selezione e la produzione dei materiali di lettura in inglese per la scuola media”, in *Educazione alla lettura. Atti del convegno LEND*, Vol. 2, Zanichelli, Bologna, pp. 124-134.

(1983) “Il ruolo della grammatica in lingua materna e in lingua straniera: principi per una collaborazione interdisciplinare nella scuola media inferiore”, in M.V. Matarese Perazzo (ed.), *Insegnare la lingua. Interdisciplinarità L1-L2*”, Bruno Mondadori, Milano, pp. 62-70.

(1983) “La formazione iniziale degli insegnanti di lingua e letteratura straniera”, in L. Panzeri Dosaggio (ed.), *Insegnare la lingua. Formazione e aggiornamento*, Bruno Mondadori, Milano, pp. 70-80.

(1984) “From ‘one English’ to ‘many Englishes’: linguistic, cultural and teaching problems”, *Problems and Experiences in the Teaching of English* 3, pp. 3-8.

(1984) “Models for the pre-service training of teachers of English as a foreign language in Italy”, *Perspectives: A Journal for Teachers of English as a Foreign Language* VIII (3), pp. 1-11.

(1984) “Una proposta d’educazione linguistica a livello di consiglio di classe nella fase iniziale della I media”, in *L’educazione linguistica dalla scuola di base al biennio della superiore. Atti del convegno CIDI-LEND*, Vol. 2, Edizioni Scolastiche Bruno Mondatori, Milano, pp. 142-151.

(1986) “Contrastive analysis and the EFL teacher in Italy”, *Problems and Experiences in the Teaching of English* 4, pp. 22-28.

(1986) “The EFL teacher as a cross-cultural interpreter”, *Perspectives: A Journal for Teachers of English as a Foreign Language* XI (1), pp. 60-67.

(1987) “Grammatiche dell’inglese per insegnanti e studenti”, in P. Ciavatta, G. Centazzo and M. Currò (eds), *Grammatica e insegnamento comunicativo*, Bruno Mondadori, Milano, pp. 100-111.

(1987) "La scelta della L2 nella scuola dell'obbligo: monolinguismo o plurilinguismo?", *LEND. Lingua e Nuova Didattica* XVI (2), pp. 46-53.

(1987) "The teacher of English as classroom researcher: some guidelines and practical advice", *Perspectives: A Journal for Teachers of English as a Foreign Language* XII (1), pp. 41-49.

(1987) "What should EFL teachers know about lexis?", in R. Boardman and S. Holden (eds), *English in School: An Overview*, Modern English Publications, Oxford, pp. 26-32.

(1988) "Dai modelli teorici alle grammatiche pedagogiche", *Le lingue nel mondo* LIII (1-2), pp. 25-29.

(1988) "Il centro linguistico e audiovisivi universitario (CLAU) dell'Università di Torino", in A. Ciliberti (ed.), *I centri linguistici nelle università italiane: prospettive e confronti*, Quaderni CIAL, Trento, pp. 63-72.

(1988) "L'insegnamento delle lingue nell'università: problemi e prospettive", in A. Ciliberti (ed.), *I centri linguistici nelle università italiane: prospettive e confronti*, Quaderni CIAL, Trento, pp. 21-36.

(1989) "Creativity and productivity in English lexis", in D. Hill and S Holden (eds), *Creativity in Language Teaching*, Modern English Publications, Oxford, pp. 80-85.

(1989) "Il '*who is who*' dei dizionari di inglese: loro tipologia, tradizione e rilevanza per gli studenti", in M.T. Prat Zagrebelsky (ed.), *Dal dizionario ai dizionari. Orientamento e guida all'uso per studenti di lingua inglese*, Tirrenia Stampatori, Torino, pp. 13-86.

(1989) "Insegnamento e ricerca: due mondi separati o complementari? Alcuni spunti di riflessione per gli insegnanti di lingue e letterature straniere", *Laboratorio degli Studi Linguistici* 2, pp. 27-40.

(1990) "You want to find out more about your students' learning strategies: why not ask them?", in D. Hill and S. Holden (eds), *Effective Teaching and Learning*, Modern English Publications, Oxford, pp. 75-79.

(1991) "Gli insegnanti di lingua straniera e di italiano: conoscersi, parlarsi, cooperare", in C. Marello and G. Mondelli (eds), *Riflettere sulla lingua*, La Nuova Italia, Firenze, pp.53-65.

(1991) "The study of English Language in Italian universities: a personal review", in M.T. Prat Zagrebelsky (ed.), *The Study of English Language in Italian Universities*, Edizioni dell'Orso, Alessandria, pp. 3-28.

(1992) "In quanti modi si può morire nell'inglese contemporaneo", in R. Rutelli and A. Johnson (eds), *I linguaggi della passione*, Campanotto Editore, Udine, pp. 361-366.

(1992) "Processes of lexical and semantic innovation in contemporary English: The case of the Gulf War", *Textus* V, pp. 111-122.

(1993) "L'innovazione lessicale nell'inglese contemporaneo: tipologie descrittive e teorie esplicative", in V. De Scarpis, L. Innocenti, F. Marucci and A. Pajalich (eds), *Intrecci e contaminazioni*, Supernova, Lido (Venezia), pp. 481-490.

(1993) "Processes of lexical innovation and neologisms in contemporary English: the case of Great Britain and European integration", in D. Hart (ed.), *Aspects of English and Italian Lexicology and Lexicography. Papers Read at the Third National Conference of History of the English Language*, Bagatto Libri, Roma, pp. 30-38.

(1994) "Blending in English: some new examples from the 80's", in R. Bacchielli (ed.), *Historical English Word-formation. Papers Read at the Sixth National Conference of the History of English*, Quattro Venti, Urbino, pp. 219-225.

(1994) "Dal laboratorio linguistico al centro linguistico: dalla dipendenza all'autonomia" in L. Mariani (ed.), *L'autonomia nell'apprendimento linguistico*, La Nuova Italia, Firenze, pp. 189-198.

(1994) "L'apprendimento 'autonomo' nei centri linguistici", in G. Bernini and M. Pavesi (eds), *Lingue straniere e università. Aspettative e organizzazione didattica*, Franco Angeli, Milano, pp. 157-165.

(1994) "L'insegnante della scuola superiore come 'consigliere per l'apprendimento'", in D. Corno and M.G. Dandini (eds), *La voglia di insegnare,* Regione Piemonte. Assessorato Istruzione, Torino, pp. 64-70.

(1994) "Lo studio dell'innovazione lessicale nell'inglese contemporaneo: ipotesi e stili di ricerca a confronto", in H. Pessina Longo (ed.), *Atti del seminario internazionale di studi sul lessico*, CLUEB, Bologna, pp. 131-138.

(1994) (with V. Pulcini) "L'inglese contemporaneo: aspetti descrittivi e forme di apprendimento", in F. Marenco (ed.), *Guida allo studio della lingua e della letteratura inglese*, Il Mulino, Bologna, pp. 43-64.

(1995) "Lo studio di due lingue straniere nella scuola media: verso l'Europa", in *Atti del seminario* '*Lingue a incastro*', Ufficio Lingue Straniere dell'Assessorato della Pubblica Istruzione della Valle d'Aosta, Aosta, pp. 69-75.

(1995) "Self-access materials in the learning of English as a foreign language: human and technological resources", in G.C. Cecioni and C. Cheselka (eds), *Proceedings of the Symposium on Language and Technology*, Università degli Studi di Firenze, Centro Linguistico d'Ateneo, CUSL, Firenze, pp. 435-439.

(1995) "Travaux récents sur l'anglais dans le monde", *Liber*, 23 Juin 1995, pp. 14-15.

(1995) "Watching English lexis change: from dictionaries of neologisms to computerised corpora", *Textus* VIII (2), pp. 249-265.

(1997) "Le lingue straniere nella scuola italiana degli anni '90" in V. Pulcini (ed.), *La didattica della lingua inglese. Percorsi per l'aggiornamento 1994-1995*, Edizioni dell'Orso, Alessandria, pp. 57-68.

(1998) "A centaur and/or a galaxy? Images for English today", in C. Taylor Torsello, L. Haarman and L. Gavioli (eds), *British/American Variation in Language, Theory and Methodology*, CLUEB, Bologna, pp. 259-267.

(1998) "Discussione dei risultati del questionario *Le lingue straniere all'università* rivolto ai laureandi di Pavia", in M. Pavesi and G. Bernini (eds), *L'Apprendimento linguistico all'università: Le lingue speciali*, Bulzoni Editore, Roma, pp. 109-122.

(1998) "Il lessico: descrizione, insegnamento, apprendimento", in M.T. Prat Zagrebelsky (ed.), *Lessico e apprendimento linguistico. Nuove tendenze della ricerca e pratiche didattiche*, La Nuova Italia, Firenze, pp. 1-80.

(1998) "La galassia dell'inglese oggi, tra stabilità, variazione e cambiamento", in P. Bayley and F. San Vicente (eds), *In una Europa plurilingue. Culture in transizione*, CLUEB, Bologna, pp. 3-8.

(1998) "The ICLE project: the Italian subcorpus of learner English on computer", *SLIN Newsletter* 19, pp. 18-23.

(1999) "Il metalinguaggio per descrivere la lingua inglese: un lessico specialistico ad hoc o condiviso da altre lingue?", in G. Azzaro and M. Ulrych (eds), *Transiti linguistici e culturali. Atti del XVIII congresso nazionale dell'Associazione Italiana di Anglistica*, E.U.T., Trieste, pp. 25-38.

(1999) "Watching English expressions – and genres – enter Italian. 'Briefing' and 'question time'", *Anglistica. Annali Istituto Orientale Napoli* III (1), pp. 107-119.

(2001) "La valutazione dei livelli d'ingresso all'università", in F. Gattullo (ed.), *La valutazione degli apprendimenti linguistici*, La Nuova Italia, Firenze, pp. 267-282.

(2001) "L'uso dei *corpora* linguistici nella descrizione e nell'apprendimento delle lingue straniere con particolare riferimento all'inglese", in V. Pulcini (ed.), *La didattica della lingua inglese. Percorsi per l'aggiornamento 1996-1999*, Edizioni dell'Orso, Alessandria, pp. 123-138.

(2001) "L'uso dei corpora nell'analisi contrastiva di saggi argomentativi di studenti universitari italiani e anglosassoni: *I think* versus *I feel*", in G.L. Beccaria and C. Marello (eds), *La parola al testo. Scritti per Bice Mortara Garavelli*, Edizioni dell'Orso, Alessandria, pp. 331-339.

(2001) "The Longman Grammar of Spoken and Written English: lexico-grammatical patterns, multi-word lexical units, idiomatic phrases, collocations, inserts, binomials, lexical bundles… and other strange things", *Studi Italiani di Linguistica Teorica e Applicata (SILTA)* XXX (2), pp. 327-334.

(2001) Review of: *Learner English on Computer*, edited by Sylviane Granger, Longman, London and New York, 1998, *International Journal of Corpus Linguistics (IJCL)* 5 (2), pp. 259-262.

(2002) "Contractions. A corpus-based analysis of argumentative essays written by English and Italian university students", in G. Iamartino, M.L. Bignami and P. Evangelisti (eds), *The Economy Principle in English: Linguistic, Literary and Cultural Perspectives. Proceedings of the XIX Conference of the Associazione Italiana di Anglistica*, UNICOPLI, Milano, pp. 242-254.

(2002) "Italian", in S. Granger, E. Dagneaux and F. Meunier (eds), *The International Corpus of Learner English, ICLE, Handbook,* with CD-ROM, UCL, Presses Universitaires de Louvain, Louvain-la-Neuve, pp. 33-35.

(2002) "The status of English in Italy", in S. Granger, E. Dagneaux and F. Meunier (eds), *The International Corpus of Learner English, ICLE, Handbook,* with CD-ROM, UCL, Presses Universitaires de Louvain, Louvain-la-Neuve, pp. 106-110.

(2002) "Una guida alle guide all'uso dei dizionari bilingui italiano-inglese /inglese-italiano", in E. Ferrario and V. Pulcini (eds), *La lessicografia bilingue tra presente e avvenire*, Edizioni Mercurio, Vercelli, pp. 139-147.

(2003) "Computer learner corpora, or how can we turn our students' interlanguage into a resource for EFL research and teaching?", *Vigo International Journal of Applied Linguistics (VIAL)* 0, pp. 103-120.

(2004) "Conclusions and outlook for the future", in M.T. Prat Zagrebelsky (ed.), *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage*, Edizioni dell'Orso, Alessandria, pp. 223-247.

(2004) "'Ethnic' rather than 'racial': a 'politically correct' or a 'scientifically correct' choice? Some corpus-based reflections", in E. Barisone, M.L. Maggioni and P. Tornaghi (eds), *The History of English and the Dynamics of Power. Proceedings of the 8th Italian Conference on the History of the English Language*, Edizioni dell'Orso, Alessandria, pp. 191-203.

(2004) "From corpus linguistics to computer learner corpora", in M.T. Prat Zagrebelsky (ed.), *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage*, Edizioni dell'Orso, Alessandria, pp. 11-60.

(2004) "I corpora nella descrizione e nella didattica delle lingue: una nuova risorsa per gli insegnanti", *LEND. Lingua e Nuova Didattica* XXXIII (1), pp. 22-37.

(2004) "Italian advanced EFL learner language: corpus-based findings", in M.T. Prat Zagrebelsky (ed.), *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage*, Edizioni dell'Orso, Alessandria, pp.61-115.

(2006). "Studiare l'inglese in inglese: resoconto di alcune esperienze alla Facoltà di Lingue e Letterature Straniere di Torino", *Rassegna italiana di linguistica applicata (RILA)* XXXVIII (1), pp. 121-132.

(2007) "L'introduzione della corpus linguistics o linguistica dei corpora nelle Università italiane: una ricostruzione personale dagli anni '70 a oggi", in N. Minerva (ed.), *Lessicologia e lessicografia nella storia degli insegnamenti linguistici*, Vol. 2, CLUEB, Bologna, pp. 203-213.

(2007) "Lexico-grammatical errors in Italian EFL university students' written productions: a corpus-based project", *English Studies 2006*, pp. 171-181.

1. Corpora and Discourse Analysis

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 37-53 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Weakness and fear: a fragment of corpus-assisted discourse analysis

Paul Bayley – University of Bologna

1. Introduction

This paper will present an analysis based on a collection of political texts and will attempt to combine two approaches to the study of language, Corpus Linguistics (see McEnery and Wilson 2001) and Discourse (or Text) Analysis (see, for example Van Dijk 1997), which have been considered by some, especially among the mainstream members of both schools, as incompatible. Although both approaches are based on the analysis of naturally occurring instances of language use, there are many differences between them. For example, corpus linguistics has traditionally dealt with instances of language which have been stripped of their context, while discourse analysis sees as fundamental the context in which utterances or stretches of discourse occur. Corpus linguists typically deal with small units of meaning, with the 'node' at their centre, while discourse analysts posit the text as the basic unit of meaning. Discourse analysis is based on subjective interpretation of texts and is often politically committed while corpus linguistics, on the other hand, makes claims for its objectivity and holds that its analyses are replicable.

From some points of view, both approaches have the weaknesses of their strengths. Discourse analysis has proven its ability to furnish detailed analysis of individual texts, or relatively short stretches of text, but its emphasis on close reading limits the amount of data that can be analysed (but see Prat Zagrebelksy 1992 for an analysis of a large collection of texts without the aid of the computer). Corpus linguistics has provided us with the tools to look at huge amounts of naturally occurring linguistic data, but, especially with very large corpora such as the Bank of English and the BNC, very little information about context is available. Distinctions are frequently made between quantitative studies (of corpora) and qualitative studies (of individual texts), and discourse analysts have argued in favour of the latter because of, *inter alia*, the "enormous influence a single text may have" (Edwards and Martin 2004: 147-8). To this I would answer that it is not easy to evaluate the importance of a single text unless it can seen against the backdrop of other texts. With the growing use of relatively small specialised corpora in what has been called the third phase of corpus linguistics (Prat Zagrebelsky 2004, 20-23), I will argue that it is possible to examine a corpus in order to identify key meanings across a corpus and then to shift back and forth from the corpus data to the textual dimension (see, for example, Partington *et al*. 2003; Miller 2006).

However, conducting Corpus-assisted Discourse Analysis is an extremely time consuming business and it takes up a great deal of (editorial) space. If a discourse analysis of a single text, or a corpus linguistics analysis of the patterns associated with a relatively infrequent lexeme might reasonably be expected to unfold within the space of a research article, an attempt to combine an analysis of a set of texts using the tools of both corpus linguistics and discourse analysis would require at the very least a book-length presentation; the quantity of material and of analytical categories that need to be taken into consideration is enormous. It would be necessary to describe the construction of the corpus, reconstruct the contextual configurations of the text types that compose it, perform analyses at diverse levels, shifting back and forth between the corpus data and the texts which compose it, and finally compare findings with those in large general corpora. Consequently, in a short paper such as this, many levels of analysis will have to be sacrificed. I will be able to indicate some directions which will allow us to say something useful both about phraseology and the company that words keep but also about discourse and text types.

2. The corpus

In the corpus linguistics tradition based on very large corpora, the analyst is not expected to know very much about the texts which make up the corpus. In this paper I will take a completely different approach; I will work with a small, highly specialised corpus that is composed of texts of American political speeches, in which I have a long-term interest. Many of the texts I know and had studied (and even watched as a TV spectator sport) before I even began to look at corpora (Ragazzini *et al.* 1985; Bayley and Miller 1993). It has been said that in corpus-assisted discourse studies analysts actually read their corpora. In this case I had read them before and had had my own subjective intuitions; the corpus has allowed me to verify whether what I had seen as typical patterns really do recur across a wider set of similar texts.

The corpus, organised according to diachronic principles, is composed of the transcripts of all the State of the Union addresses (which take place annually), inaugural speeches, acceptance speeches and televised debates (which take place every four years, with the exception of debates in the elections of 1964, 1968 and 1972) held between 1960 and 2004. It is composed of 98 texts and amounts to approximately 675,000 tokens, corresponding to about 100 hours of talk. The analysis was conducted with the aid of the software Wordsmith Tools 4.0, using default settings. I have previously used the corpus, comparing it to a rather different corpus of British political language (Bayley 2005, 2007).

3. Identifying patterns in the corpus

One of the intuitions that I had from a close reading of some of these speeches and debates was that one of the foundations of US political rhetoric was the construction of symbolic oppositions through the co-selection of antonymic pairs such as *WEAKNESS* and *STRENGTH* or *FEAR* and *COURAGE*, as illustrated in examples 1 and 2:

(1) This debate tonight has made crystal clear a challenge that is old as America – the choice between hope and fear, change or more of the same, the courage to move into a new tomorrow or to listen to the crowd who says things could be worse. (Clinton, televised debate, 11 October 1992).

(2) Where the world super-powers are concerned there is no acceptable alternative to peaceful negotiation. Because this will be a period of negotiations we shall restore the strength of America so that we shall always negotiate from strength but never from weakness. (Nixon, acceptance speech, 8 August 1968).

A systematically organized corpus of Presidential speeches should allow us to verify whether these are isolated meaning-making practices or whether they are part of a general discourse pattern. I shall begin with the least frequent of these words – the noun *WEAKNESS*. It occurs 33 times in the corpus[1] (a low relative frequency of 0.004 per hundred words) and is co-selected with *STRENGTH* in 9 instances (within a search window of 10 words either side of the node), as illustrated below.[2]

1 at our diversity is a **weakness**, it is our greatest **strength**

2 it a **strength**, not a **weakness**.

3 e’ve turned it into a **weakness**. Now again, the White House

4 **rength** and never from **weakness**. And as we seek through nego

5 lve, but we are given **weakness** when we need **strength**; vacil

6 can compensate for my **weakness**, and your wisdom can help to

7 arity lead to peace - **weakness** and ambivalence lead to war.

8 **ength**, because of the **weakness** of other major free world po

9 ivalence lead to war. **Weakness** tempts aggressors. **Strength**

The low frequency of *WEAKNESS* may seem to demonstrate that my intuition was quite mistaken, and to an extent this may be true. However, it does confirm another one, that is that negative symbols are dispreferred in political discourse, and the majority of instances of the term occur in adversarial discourse types (acceptance speeches and debates) and they are used to criticise an opponent. Moreover, if you look at the fourth and last lines in the concordances above, it is possible to discern a particular semantic pattern which may be glossed ‘peace through strength’, which is a common motif for foreign policy (see example 2, above).

[1] The absolute frequency of the lexeme *WEAK*, including the adjective form and the verb *WEAKEN* is 118.

[2] Because I have chosen a search window which is wider than that traditionally used in corpus linguistics, not all of the co-selections appear in the figure.

STRENGTH is a far more frequent noun than *WEAKNESS* and the corpus yields 313 instances (relative frequency, 0.04), co-selected with *PEACE* in 38 occasions, calculated with a collocate horizon of ten words to the left and right of the node. The two words also have a strong mutual attraction over longer segments of text; doubling the collocate horizon yields 56 co-selections, and a total of 81 instances of *STRENGTH*. Moreover, the most frequent three-word cluster (excluding "the strength of" and "strength of the") is "peace through strength", which is selected in the discourse of Presidents from Carter to G. W. Bush (excluding Reagan and Clinton). All but one of the 38 instances of co-selection realise a meaning which can be glossed as 'the stronger America is, the greater is the possibility of peace'. The exception is in example 3:

(3) Weapons do not make peace. Men make peace. And peace comes not through strength alone, but through wisdom and patience and restraint (Johnson, acceptance speech, 27 August 1964).

Twenty examples of co-selection are given below:

1 make **peace.** And **peace** comes not through **strength** alone, but

2   And to maintain that **peace** requires **strength.** America h

3 eace with freedom must also be based on **strength**- - economi

4 reeze, and we insisted on **peace** through **strength.** From Ango

5 ll continue our policy of **peace** through **strength.** I will m

6 President Reagan said no, **peace** through **strength.** It worked

7 ue **peace,** fight terrorism, increase our **strength,** renew our

8 pects America's policy of **peace** through **strength.** The Unite

9 ntry because we've got to have military **strength** to preserv

10 that we want to preserve **peace** through **strength.** We also w

11 erybody now realizes that **peace** through **strength** works, and

12 at has done exactly that. **Peace** through **strength** works.

13 ct: **Strength** in the pursuit of **peace** is no vice. Isolationi

14 le. **Strength** is imperative for **peace,** but the two must go h

15 nal **strength** and security is identical to the path to **peace**

16 hose **strength** is respected--is essential to continued **peace**

17 of **strength** because we're the guardians of the **peace.** In a

18 ary **strength** has always been maintained to keep the **peace,**

19 ose **strength** is respected--is essential to continued **peace**

20 ary **strength** of the United States to preserve the **peace.** We

While *WEAKNESS* and *STRENGTH* seem to be associated with questions of foreign policy, *FEAR* and *COURAGE* seem to belong, generally but not exclusively, to the domain of domestic politics. The lexeme *FEAR* occurs in the corpus 102 times (see Bayley 2005), 81 of which in the noun form, and is generally associated with negative attitude (often construed through its collocation with antonyms – “the choice between hope and fear” - or with negatively charged lexical items – “fear and hate”). It is co-selected with its antonym *COURAGE* only once (see example 1), however, which means that my original hypothesis, based on a close reading of some of Clinton’s speeches, is not confirmed; if anything the opposition is implicit and can be identified in longer stretches of texts. Meaning patterns fit into two distinct but related semantic motifs. The word is used in a general semantic framework according to which ‘fear is the strategy of my adversary’ (for example “all you have is fear”, “the only thing he has to offer is fear itself”) and ‘fear generates failure’ (for example, “[…] only fear prevents their reduction”, “Fear of the future was throttling […]”). Examples of these can be seen in the 20 instances selected below.

1 rats to end the politics of **fear** and save Social Security,

2 thing he has to offer is **fear** itself. That outlook is ty

3 s nuclear program to incite **fear** and seek concessions. Amer

4 govern with negativism and **fear** of the future, but with vi

5 American people. There is a **fear** that our best years are be

6 opposition is rooted in two **fears**: first, that our trading

7 the choice between hope and **fear**, change or more of the sam

8 rcials and media events and **fear** messages and personal atta

9 question between trust and **fear**, and I would say, I think,

10 going in this campaign is **fear**. You're spending millions

11 agenda, and all you have is **fear**, that's all you can use. W

12 lems, mankind can turn from **fear** towards hope. From the tim

13 being laid off every week. **Fear** of the future was throttli

14 But we must remember that **fear** of a recession can contrib

15 turbulence and doubt, and **fear** and hate. Throughout this

16 reedom. They did not mingle **fear** and joy, in expectation th

17 t the forces of bigotry and **fear** and smear. Our problems ar

18 ivision; offering hope, not **fear** or smear. We do offer the

19 So long as fanaticism and **fear** brood over the affairs of

20 forces, the hatreds, the **fears** that divide the world.

As a verb, *FEAR* is used to construct the speaker as fearless, in linguistic contexts of negative mood ("I do not fear") and the other (the adversary) as fearful ("others fear that"):

```
1 strust the future; I do not fear what is ahead. For our pro
2 th us I say that we neither fear competition nor see it as
3  But we in America need not fear change. The values on whic
4 s. But America need have no fear. We can thrive in a world
5  Americans will never again fear the snooping, harassment,
6  Shelby, America need never fear for our future. And we see
7 he hope of it. Where others fear trade and economic growth,
8 terrorists and their allies fear and fight this progress ab
9 there will be naysayers who fear that we won't be equal to
10 erns. Some in the Pentagon fear that too much priority has
11 ll said the same thing, he fears the initiative would take
12 g further behind; why they fear the current generation wil
13 nor, and say - should they fear us? Should they welcome ou
```

The antonym of *FEAR*, *COURAGE*, occurs 97 times and thus has a similar relative frequency to fear. As I have already noted, the two terms are not co-selected. However, they are connected by two opposing semantic motifs: 'fear is a quality of my adversary' and 'courage is a virtue of America'. In fact *COURAGE* is co-selected with *AMERICA* (or *AMERICANS*) in 25 instances, 20 of which are illustrated below. The positive judgement that is conferred upon *COURAGE* is further confirmed by its frequent collocation with other nouns expressing American virtues, such as *WISDOM*, *FREEDOM*, *STRENGTH*, *FAITH*, *VISION*, *COMPASSION*, *COMMITMENT*, *HOPE*, *CONFIDENCE* and *IDEALISM*.

```
1    are made of sugar candy." We Americans have courage. Ame
2 e American people brought us back - with quiet courage and
3 ican people have responded magnificently, with courage and
4  American belief: the belief that strength and courage and
5 s America's fighting men have set a record for courage and
6 s character.  America has need of idealism and courage beca
7  America is back, looking to the eighties with courage, con
8 ed accomplishment. America at its best is also courageous.
9 rtels and other shortages. American wisdom and courage righ
10  and the civilized world. And by our will and courage, thi
```

```
11  and hard work we do today; For an America of courage whos
12 be America's new direction. Let us summon the courage to
13 ies of courage and compassion that we strive for in America
14  nity, courage, compassion and character. America, at its b
15  nt of courage, idealism, and bipartisan unity can change A
16 nd the courage of the American people. I am speaking of a n
17 nd the courage of those great Americans who met in Philadel
18 erican courage overcome American challenges. When Lewis Mor
19 to the courage, patience, and strength of our people, Ameri
20 on and courage to reinvent America. When our founders bold
```

4. Shifting back and forth from corpus to text

From these meaning patterns taken from across the corpus it is possible to move towards specific texts by identifying, for example, those in which our search words have the highest frequency, or those in which they tend to be co-selected. The software can help us with this, either through Wordsmith's plotting function, or simply by ordering the concordances in chronological order. Choosing the latter route, the first two instances of fear immediately attract attention.

```
1 us never negotiate out of fear, but let us never fear to n
2 of fear, but let us never fear to negotiate. Let both side
```

Two instances of the search word occur within the same clause-complex and they are recognisable as parts of a famous phrase uttered by J. F. Kennedy in his inaugural speech:

(4) Let us never negotiate out of fear, but let us never fear to negotiate. (Kennedy, inaugural address, 20 January 1960).

The first instance of *FEAR*, realised in an adverbial expressing manner, can be associated, as I will argue shortly, with *WEAKNESS*, while the second, the lexical unit of the verbal group, has another meaning: "fear to negotiate" could be substituted with "be apprehensive about negotiating". The two instances of the same lexeme thus realise very different meanings.

A further look at the corpus, sorted in chronological order, reveals that within the same text segment *WEAKNESS* also co-occurs twice:

1 dare not tempt them with **weakness.** For only when our arms a
2 ivility is not a sign of **weakness,** and sincerity is always

This matching of two occurrences of two negatively charged words suggests that the text that they belong to may be of interest. To make a telling analysis of these occurrences, it would be necessary to take into consideration, at the very least, the social and historical context in which the speech was made, beginning from a consideration of Cold War discourse, the context of situation in which the speech was made, the co-text to which the example belongs, and its intertextual environment. However, I shall merely indicate in a cursory fashion a few ways in which a corpus investigation can interact with discourse analysis, beginning firstly with some features of the context of situation according to the categories of field (what is going on), tenor (who is taking part) and mode (the means of communication) (Halliday 1978), which, according to systemic functional linguistics contribute to the determination of lexicogrammatical patterns in diatypical language varieties.

As far as the first category is concerned, the occasion was a solemn and sober ritual held in front of Capitol Hill in Washington DC (very different from an acceptance speech); the speech immediately followed the declaration of the oath prescribed the US Constitution. Of the text types in the corpus, the inaugural is the least ‘adversarial’ in terms of domestic politics, and we would expect little or no negative construal of political opponents. The subject matter or topic is typically a symbolic representation of the tasks that await the administration over the following four years, and thus varies according to historical circumstances and the priorities of the new President. Thus, while F. D. Roosevelt’s first inaugural was grounded in the depression, Kennedy’s situated itself in the midst of the Cold War, but also, in his words “a struggle against the common enemies of man: tyranny, poverty, disease and war itself”.

Concerning tenor, the inaugural speech is characterised by a speaker who holds both discursive power (he delivers a monologue) and socio-political power (he is no longer seeking election but celebrating his victory). The addressee is multiple; a primary addressee – the authorities who sit behind him (some of whom are directly addressed at the opening - Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy), the audience in front of him and in general the nation at large (fellow citizens). But because he is the President of a super-power, the nations of the world are a tertiary addressee, and as we shall see, Kennedy explicitly constructs them as such. We would thus expect the speaker to enact the role of leader both through solidarity and unity ("We observe today not a victory of party but a celebration of freedom") as well as power, using imperative mood ("ask not what your country can do for you; ask what you can do for your country").

The mode of discourse may be glossed as a written-to-be-spoken monologue, read from a text and not extemporised, directed to a live audience, to a television and radio audience, to the newspapers which will reproduce it, and ultimately to posterity. Inaugurals are typically quite short (though Kennedy's was among the shortest – 1,365 tokens). Diction can be expected to be deliberate and firm (Kennedy's speech lasted for 14 minutes and was thus uttered at a very slow 100 words per minute). The speech will also draw upon and contribute to the tradition of US political rhetoric and so we would expect to find 'poetic devices' such as grammatical parallelism, repetition and alliteration (see Miller 1993).

Let's begin to look at one segment of the text. The body of the text is marked by a series of repetitions ("to those who […] we pledge") functioning as hyperthemes and constructing a series of secondary addressees, actors in the field of international relations – European nations, newly independent nations, poor nations, Latin American Nations, the United Nations, and the Warsaw Pact nations – taking up just over half of the text.

(5) To those old allies whose cultural and spiritual origins we share, we pledge [...]
To those new states whom we welcome to the ranks of the free, we pledge [...]
To those peoples in the huts and villages of half the globe struggling to break the bonds of mass misery, we pledge [...]

To our sister republics south of our border, we offer a special pledge [...]
To that world assembly of sovereign states, the United Nations, our last best hope in an age where the instruments of war have far outpaced the instruments of peace, we renew our pledge [...]
Finally, to those nations who would make themselves our adversary, we offer not a pledge but a request [...].

This form of what may be called ‘indirect apostrophe’ is a consolidated feature of the American rhetorical tradition. Grammatically, the structure turns around the phrase “to those” following by a relative clause or classifier (“to those who care”, “to those new States”), and is either preceded or followed by a verb of saying in the first person (“I say to those”, “to those […] I say”). This is quite a complex configuration of lexicogrammatical features and yet the corpus yields 45 instances, 20 of which are illustrated below:

```
1  odds and split asunder. To those new states whom we welc
2 dignity for all workers. To those who care for our sick,
3 interests of the nation. To those who have stood with me
4  to follow their dreams. To those imprisoned in regimes h
5 t of abortion on demand. To those who say this violates a
6 ng-term care. (Applause) To those who would cut Medicare
7 do not now have freedom. To those neighbors and allies wh
8 extends its hand to you. To those nation-states that wish
9 m fear in the world. And to those who say that law and or
10 future. And I would say to those I know there are more y
11   for your support. And to those who have not felt able
12 the history of mankind. To those who have sought to divi
13 e in the last few days, to those who say the progress of
14  writ may run. Finally, to those nations who would make
15 we know it. We will say to those on welfare: you will ha
16  me in spirit in saying to those who would malign our st
17 ness and greatness. And to those timid souls, I repeat t
18 dren. (Applause.) I say to those who are on welfare, and
19 nly helped to unite us. To those who would provoke us we
20 f the Democratic Party. To those millions who have been
```

The clause-complex we have identified through concordance procedures belongs to the final section, addressed to “those nations who would make themselves our adversary", in which Kennedy makes explicit the speech act which he will perform – a request. The complete text segment, divided into clauses, is as follows:

(6) (a) Finally, to those nations who would make themselves our adversary, we offer not a pledge but a request:
(b) that both sides begin anew the quest for peace,
(c) before the dark powers of destruction unleashed by science engulf all humanity in planned or accidental self-destruction.
(d) We dare not tempt them with weakness.
(e) For only when our arms are sufficient beyond doubt can we be certain beyond doubt
(f) that they will never be employed.
(g) But neither can two great and powerful groups of nations take comfort from our present course –
(h) both sides overburdened by the cost of modern weapons, both rightly alarmed by the steady spread of the deadly atom, yet both racing to alter that uncertain balance of terror that stays the hand of mankind's final war.
(i) So let us begin anew,
(j) remembering on both sides that civility is not a sign of weakness,
(k) and sincerity is always subject to proof.
(l) **Let us never negotiate out of fear**,
(m) **but let us never fear to negotiate**.
(n) Let both sides explore what problems unite us
(o) instead of belaboring those problems which divide us.
(p) Let both sides, for the first time, formulate serious and precise proposals for the inspection and control of arms,
(q) and bring the absolute power to destroy other nations under the absolute control of all nations.
(r) Let both sides seek to invoke the wonders of science instead of its terrors.
(s) Together let us explore the stars, conquer the deserts, eradicate disease, tap the ocean depths and encourage the arts and commerce.
(t) Let both sides unite to heed in all corners of the earth the command of Isaiah to "undo the heavy burdens ... [and] let the oppressed go free."
(u) And if a beachhead of co-operation may push back the jungle of suspicion,

> (v) let both sides join in creating a new endeavor, not a new balance of power, but a new world of law,
> (w) where the strong are just and the weak secure and the peace preserved.

Having identified a richer co-text, our analysis may take a number of directions. The first could be that of the meanings that are constructed through lexical relations. Eight clauses prior to clauses (l) and (m), Kennedy had said:

(7) We dare not tempt them with **weakness**. For only when our arms are sufficient beyond doubt can we be certain that they will never be employed.

WEAKNESS is again repeated in clause (j), just before the *FEAR* clause complex, in a negative mood environment ("civility is not a sign of weakness") in which the noun *CIVILITY* can be seen as anticipating the verb *NEGOTIATE*. Because of the grammatical and semantic parallelism of clauses (d) and (l) ("we dare not tempt them with weakness" and "let us never negotiate out of fear") this seems to establish a cohesive tie between two lexical items, *WEAKNESS* and *FEAR***:** they are construed as being a member of the same class of word, co-hyponyms of "human or institutional failings". Kennedy also makes the nature of this *WEAKNESS* explicit – "only when our arms are sufficient beyond doubt" – thus contributing to the extended 'peace-through-strength' motif which was illustrated in example 2 above.

Another grammatical feature, or rather system, that often proves interesting for discourse analysis is transitivity, particularly within the field of critical discourse analysis (see, for example, Fowler 1996; Fairclough and Wodak 1997). The first half of this passage seems to abound in problems related to grammatical agency. In identifying his addressee, Kennedy uses the phrase "those nations who would make themselves our adversary" and by so doing exonerates the USA from any responsibility for the Cold War, laying the blame for the nuclear arms race directly at the feet of the USSR, whose own inclination ("would make themselves") has created the international tension.

In a similar vein it could be observed that in the non-finite clause (h), Kennedy emphasises the economic rather than the social and humanitarian aspects of the arms race (“overburdened by the cost of modern weapons”) and at the same time he selects a grammatical agent, ‘costs’, which severs the relationship between ‘spender’ and the ‘process of spending’. Moreover, he selects the adjective *OVERBURDENED*, which, looking at the data in the BNC, generally associates the ‘attribute’ with an unwilling or involuntary ‘carrier’. If, instead, he had said “Both sides, overspending on modern weapons”, he would have realised a very different meaning. Similarly, he represents nuclear weapons as self-propagating rather than the product of intense human research and labour over more than a decade (“the steady spread of the deadly atom”). According to BNC data, the content word collocating most frequently with *SPREAD* (with the exception of *HANDS*) is *DISEASE*, followed closely by *INFECTION* – which are not typically spread voluntarily. Kennedy represents a process as a thing which alarms its real perpetrators, and which has created a monster (“that uncertain balance of terror”) that, however, is represented as contingently beneficial because it “stays the hand of mankind’s final war”, by which the President constructs the act of war as an independent actor, whose “hands” may be stayed. In the immediate context of other Presidential speeches, this may seem peculiar; just three days before Kennedy’s inauguration, his predecessor, Eisenhower had, in his farewell speech, placed the responsibility for unwarranted escalation on the “military-industrial complex”:

(8) In the councils of government, we must guard against the acquisition of unwarranted influence, whether sought or unsought, by the military industrial complex. The potential for the disastrous rise of misplaced power exists and will persist. (Eisenhower, farewell speech, 17 January 1961)

An alternative approach to the text would be to follow Martin’s appeal (1999: 38) to focus on discourse “that inspires, heartens”, and not just on what we dislike and conduct what Martin has termed “positive discourse analysis” (see Martin and Rose 2003). For example, if we go back to the text fragment above, it is evident that Kennedy uses the classical We/They dichotomy to represent an enemy, as in clause (d), “We dare not tempt them with weakness”. However this construal subtly shifts, perhaps in clause (i) and certainly in the two clauses which first attracted our attention (l) and (m); the collaborative negative imperative (“let us never negotiate”) can be interpreted as embracing both the USA and the USSR. This becomes even clearer in clause (n) that immediately follows – “Let both sides explore what problems unite us”. And thus Kennedy shifts from adversarial discourse to the discourse of mutual cooperation, further underlined by the opposition between “those problems what unite us” to the negatively construed “those problems which divide us”. In clause (q) he launches an appeal to multilateralism and finally, in the last two clauses, he undermines the presupposition underlying clause (g), that “the balance of terror” is necessary, with another appeal to multilateralism – “not a new balance of power, but a new world of law”. In clause (w) he reconstructs, with the word form *WEAK*, the meaning of *WEAKNESS* construed in clauses (e)-(m), foreseeing a world where “the strong are just and the weak secure and the peace preserved”, which represents *WEAKNESS* not as a individual or institutional failing but as a social condition to which the *STRONG* must respond with solidarity.

5. Conclusion

The objective of this paper was to give a brief illustration of how procedures of discourse analysis can be usefully combined with the procedures of corpus linguistics using specialised corpora. The brief analytical example, however, raises another issue, the contrast between positive and critical discourse analysis. What clearly emerges from the last two parts of the previous section, one presenting a critical analysis and one a positive analysis, is not only that meanings in texts may be very complex, but also that the position of the analyst is central. For example, in the critical analysis, the criticism was underpinned by the notion that Kennedy was disguising the US’s responsibility for what he called the “balance of terror”, and subsequently the “balance of power”, and that the policy of nuclear deterrence creates a dangerous state of permanent threat. This is an opinion which I happen to hold, but it is not universally held. Many scholars of international relations are convinced that nuclear deterrence is far more effective in maintaining peace than conventional weapons (see, for example the debate in Sagan and Waltz 2003), a conviction that I am prepared at least to entertain.

Similarly, the positive analysis was underpinned by the notion that multilateralism is better than unilateralism, that cooperation, in international affairs, is better than outright conflict, another opinion that I personally hold. However, it could be persuasively argued that, on the basis of empirical evidence, the supranational organisation that has the primary function of maintaining peace in the world has not proven to be particularly effective or successful in preventing conflict. I am not claiming that ultimately the subjective position of the discourse analyst can be avoided, if anything it needs to be constantly restated, but at heart, and with some limitations, I remain an unrepentant relativist, and a priori choices on what is positive and what is negative should be suspended. Perhaps we should rather let the texts speak for themselves, and let the corpus tell us which meaning-making practices are re-iterated and which are isolated.

References

Bayley, P. (2005) "The representation and construction of fear in political discourse", *Contatti* 1, pp. 87-111.

Bayley, P. (2007) "*Terror* in political discourse: from the cold war to the unipolar world", in N. Fairclough, G. Cortese and P. Ardizzone (eds), *Discourse and Contemporary Social Change*, Peter Lang, Bern, pp. 49-71.

Bayley, P. and D.R. Miller (1993) *Texts and Contexts of the American Dream: A Social Semiotic Study of Political Language*, Pitagora, Bologna.

Edwards, J. and J.R. Martin (2004) "Introduction: approaches to tragedy", *Discourse and Society* 15 (2/3), pp. 147-154.

Fairclough, N. and R. Wodak (1997) "Critical discourse analysis", in T.A. Van Dijk (ed.), *Discourse as Social Interaction. Discourse Studies: A Multidisciplinary Introduction*, Vol. II, Sage, London, pp. 258-284.

Fowler, R. (1996) "On critical linguistics", in C.R. Caldas-Coulthard and M. Coulthard, *Text and Practices*, Routledge, London, pp. 3-14.

Halliday, M.A.K. (1978) *Language as Social Semiotic: The Social Interpretation of Language and Meaning*, Edward Arnold, London.

Halliday, M.A.K. (1994) *An Introduction to Functional Grammar*, Edward Arnold, London.

Martin, J.R. (1999) "Grace: the logogenesis of freedom", *Discourse Studies* 1 (1), pp. 31-58.

Martin, J.R. and D. Rose (2003) *Working with Discourse: Meaning beyond the Clause*, Continuum, London.

Miller, D.R. (1993) "The electoral speech as register: the discursive construction of the common ground", in P. Bayley and D.R. Miller, *Texts and Contexts of the American Dream: A Social Semiotic Study of Political Language*, Pitagora, Bologna, pp. 147-198.

Miller, D.R. (2006) "From concordance to text: appraising 'giving' in *Alma Mater* donation requests", in G. Thompson and S. Hunston (eds), *System and Corpus: Exploring Connections*, Equinox, London, pp. 248-268.

McEnery, T. and A. Wilson [1996] (2001) *Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Partington, A., J. Morley and L. Haarman (eds) (2003) *Corpora and Discourse*, Peter Lang, Bern.

Prat Zagrebelsky, M.T. (1992) "Processes of lexical and semantic innovation in contemporary English: the case of the Gulf War", *Textus* V, pp. 111-122.

Prat Zagrebelsky, M.T. (2004) "From corpus linguistics to computer learner corpora", in M.T. Prat Zagrebelsky (ed.), *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage*, Edizioni dell'Orso, Alessandria, pp. 11-60.

Ragazzini, G., D.R. Miller and P. Bayley (1985) *Campaign Language: Language, Image and Myth in the U.S. Presidential Election 1984*, Cooperativa Libraria Universitaria Editrice Bologna, Bologna.

Sagan S.D. and K.N. Waltz (2003) *The Spread of Nuclear Weapons*, Norton, New York.

Van Dijk, T.A. (ed.) (1997a) *Discourse as Structure and Process. Discourse Studies: A Multidisciplinary Introduction*, Vol. I, Sage, London.

Van Dijk, T.A. (ed.) (1997b) *Discourse as Social Interaction. Discourse Studies: A Multidisciplinary Introduction*, Vol. II, Sage, London.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 55-70 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Personal narratives in children's rights discourse

Giuseppina Cortese – University of Turin

One day my father was shot. (…) My mother didn't love me, she would beat me for nothing. (http://www.newint.org/issue377/dolgion.htm)

My papa is a monster. My younger sister was beat to death. He beat me everyday, slashed me, or booted me out of the house to stand outside in the winter. Once, he let me stand beside the icy hole of the cold water, and then only one kick, I was thrown out into the hole (…) I won't go back. I will die. (http://web.syr.edu/~yli22/intro_page_1.html)

When I was just a little boy my mother broke up with my father. While living with my mother, she always beat me for no reason whatsoever. Sometimes she would strip me down to bare skin and beat me with (…) this thing they call 'rigoise', which is made with dried cow hide. After she would beat me, she goes and gets some lemons to pass all over my bruised body to ease the swelling. The stripes on me would appear like burn marks. She didn't want the neighbours to know what she was doing to me in the house. She told me the best thing I would do is leave her house. (http://quicksitemaker.com/members/immunenation/StreetkidGallery.html)

The police also encourage us to be pimped. They get money (…) There are cases when the police have a desire for a certain girl in the area. They will catch the girl and put her in prison so that the girl will have no choice but to have sex with them. (www.newint.org/issue 377/- Jessa.htm)

Children's rights? We have nothing. We're just like human garbage. Nobody needs us. Anyone can beat us. The police (…) take us to the police station where they beat us. They force us to sit (…) and beat us (…) or they (…) insert a wooden pole between the legs, right below the crotch and start rolling it. (www.newint.org/issue 377/- Dolgion.htm)

The police are very violent with us; they don't help us children, not even a little bit. Once when I was very hungry and desperate (…) I robbed a lady. A policeman saw me. He grabbed me and beat me, no-one stopped him. Then I was put in prison for nine months (…) The prison was horrible. Horrible. The police beat me; they sprayed teargas or pepper spray in my eyes regularly. I was alone but at least I got 4 meals a day. That was great. (…) The only thing in the world I hate is the police. (www.newint.org/issue 377/ricardo.htm)

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

1. Introduction[*]

The present paper is concerned with the discourse of children's rights, notably the case of street children – a particularly vulnerable group among the world's children.[1] Since the literature on child streetism is vast and the definitions provided for the condition of 'street child' are numerous and mostly imprecise, even misguiding (see Glauser 1997), I shall for the present purposes refrain from using highly debated notions such as those of 'children *in* the street' and 'children *of* the street' and I shall stick to an intuitive and comprehensive notion of 'street children' as follows: minors of any age who, for different reasons, take to the streets and dwell in public places for long periods of time. Instead, I would like to clarify that my present notion of children does not equate biological immaturity with the status of passive and 'incomplete' beings: even the qualifier 'vulnerable', so often used with regard to homeless youth, does not necessarily allude to a lack of coping skills. In consonance with constructivist paradigms, children are considered here as social actors – actors, however, whose agency and accomplishments in the social sphere are vastly marginalized and underrated.

Through their social interactions, children form their own interpretive frames to evaluate social meaning – they construct for themselves a linguistic 'sense of place' (Bourdieu 1991: 82) within their specific culture. However these discursive acts through which children construe specific ascriptions and self ascriptions within a specific setting and time remain largely untapped. Even though recent research in the sociology of childhood has largely acknowledged the fact that children have been a silenced group (Prout and James 1997: 7), much of the policy-making and professional work involving children is still informed by the concept of need, as configured by adult views (Woodhead 1997) rather than by children's voices.

[*] MIUR Cofin - Protocol N. 2005109911-002. All websites accessed in February, March and October 2007.

[1] "The world's children" is a notion which confirms the growing concern for children and children's rights – it does not, at least here, in any way assume a globalized single conception of childhood. On hegemonic European middle class models and mores concerning childhood, see Boyd 1997.

This paper will focus on street children's narrative accounts of neglect, abuse and violence – that is to say their deprivation of basic human rights. I have thus far examined the professional literature concerning street children in a number of studies of which one is briefly summarized below. Here, I intend to examine recent web-distributed texts in order to observe whether children's narratives confirm a number of polarizations highlighted by professional texts with regard to issues of violence and abuse. The main question, in other words, will be "Whom do these children systematically perceive and identify as prior sources of injustice and maltreatment?".

First of all, some background is provided on the study of children in the light of children's rights discourse.

2. Background

Human rights doctrines have been and are widely debated on ethical, ideological and legal grounds. At the heart of the discussion, which has sparked many schools of thought, lies the duality between universalist aspirations and the need to respect cultural difference. The overarching, unresolved question whether human rights norms pre-empt local practices leaves open an interpretative space which is textually embodied in the vagueness and indeterminacy of many international human rights treaties. The 1989 UN *Convention on the Rights of the Child* (henceforth CRC), which is the most relevant international agreement on children's rights, is no exception to this generalizing treatment of human subjects for the sake of universal standards.

However, the context-free, objectifying language of the CRC is now fleshed out by local experience. A transnational discourse community has been developing across a range of professions, with the important task of implementing CRC "mental software" in the most disparate legal and cultural contexts. From fieldwork case studies to UN special reports, professional and institutional discourse represents social actors – the children, the families, the state and often the security forces – who are caught between a universalist ethics of care and local standards of indifference.

In a previous study (Cortese, 2007) I focussed on the discourse of children's rights in the practices of those who are professionally operating for its dissemination and actualization worldwide. Starting from the assumption that text and talk are no mere 'reflection' of social reality, I observed how ideologies inherent in the CRC are remoulded in the arena of the implementation of children's rights. The corpus for the study consisted of reflective and/or programmatic accounts with persuasive aims such as acquiring a positive 'corporate' image, obtaining funds, struggling for attention and support in the political sphere, fighting powerful local scripts that legitimate and naturalize violence against children, winning over public opinion internationally. Hence, the notion of genre governing the corpus accommodates diverse texts in the social process of developing, maintaining and manifesting professional identity in a community of practice devoted to children's rights.

In these reports and case histories by legal/social workers and political activists negotiating CRC guidelines from within local sociocultural models of childhood and conditions of socioeconomic (in)justice, a tension can be envisaged between 'protectionist' and 'liberationist' ideology, which originates stark contrasts or changes in the ways of constructing and organizing childcare. Confronted with the experience of children who are situated socially, cognitively and emotionally in very different conditions from their Western middle class peers, professionals involved in childcare who share the axiological stance of the CRC doctrine seem to be of two minds in addressing children's moral, legal and socioeconomic standards. On the one hand they emphasize the children's predicament as victims in need of protection, on the other – and increasingly so – they stress the children's resilience and resourcefulness, highlighting their survival strategies. This confirms the findings of a previous study (Cortese 2004), that international advocacy is increasingly oriented towards a representation of homeless children dwelling in the streets as capable members in subcultures which do not share the norms and the lifestyle patterns of middle-class youth.[2] Common to the

[2] On victimized/victimizing perspectives leading to ineffective, 'palliative' approaches: "Since public attention has been drawn to the world-wide problem of street children, international organizations, government agencies, and private welfare organizations and associations have increasingly chosen them as a focus of

representation of these subcultures, and relevant to my present discussion, is the emphasis on the children's strategies in coping against adult violence and abuse. Though research and field reports focus on widely different geopolitical contexts, and in spite of the fact that street children everywhere are all but homogeneous social formations, dysfunctional families[3] and police officers[4] are commonly constructed in the literature 'about' street children as their major antagonists – notably, the perpetrators of abuse, brutality, exploitation and violence. Hence, the present discussion of texts 'by' children, aimed at giving children voice, will focus on the children's representations of those institutions – the family and the security forces – which, in most cultures and local circumstances, should be the very guardians of their rights. Voice will be pursued through the computer-assisted and manual study of lexical collocations,

their work. On the one hand this is positive; on the other hand, it sometimes leads to the 'Calcutta Syndrome', where compassion is temporarily lavished mainly on smaller children. Street children are actually wooed by many projects. As a result, the streets become particularly attractive to children from the slums. They move rapidly from project to project, taking advantage of what is on offer". (Edges Magazine 27 [2001], www.users.globalnet.co.uk, on 4/10/2007.)

[3] The following research report from South Africa identifies factors in family disruption which well apply to a variety of contexts: "the majority [of street children] leave as a result of socioeconomic and other factors within the family […]. These family factors may include: abuse of alcohol and drugs; financial problems and poverty; family violence and family breakup; poor family relationships; parental unemployment and resulting stress; physical and/or sexual abuse of children; parents absent from home as a result of personal or financial reasons (e.g. a migrant labor system); collapse of family structure; collapse of extended family; and emergence of vulnerable nuclear families in urban areas. (Le Roux 1996. www.pangaea.org/street_children/africa/safrica2.htm, on 4/10/2007).

[4] See, e.g. Human Rights Watch: "Street children risk violence at the hands of the authorities much more frequently than other children. Children on the street are beaten, tortured, sexually assaulted, and sometimes killed. Several factors contribute to this phenomenon: police perceptions of street children as vagrants and criminals, widespread corruption and a culture of police violence, the inadequacy and non-implementation of legal safeguards, and the level of impunity that officials enjoy. Street children are easy targets because they are young, often small, poor, ignorant of their rights, and frequently do not have responsible adults to look out for them. Police also have financial incentives to resort to violence against children. They beat children for their money or demand payment for protection, to avoid false charges, or for release from (often illegal) custody". (www.hrw.org/reports/2001/children/5.htm, on 4/10/2007)

colligations and text-type configurations within a framework of language as social action.

3. Narrative accounts of personal experience

The corpus for the present study consists of narrative accounts provided by street children in disparate cultural situations.[5] This of course begs a number of questions concerning first of all the peculiar nature of these personal narratives. In the first place, these web-distributed documents are English translations from native languages; further, they are obviously elicited during adult-child conversations, but the discoursal role(s) of interviewer, facilitator and immediate recipient, which are normally part and parcel of the analysis of conversational narratives and of their actual performance-in-the-telling, has/have been quite regularly removed for the sake of providing 'objective' records. If then some sources warn the reader that materials have not been tampered with by adult editing, in other cases the actual story has been cut to what the adult researcher perceived its main 'point' to be. Hence, conceding that the translation has been provided by adults who have been members of the local community long enough to have developed a good command of the local idiom and who are quite rigorous in their purpose not to interfere with the format of the story, one is faced here in any case with 'monologues'[6] covering different narrative genres, from the nearly complete structure in Labovian terms (Labov and Waletsky 1967; Labov 1972) to various narrative formations (Cortese 1995: 36 and ff.) including story capsules which focus either on the concise rendition of an event or on the

[5] The small corpus used for the present study was collected and organised under my supervision (E. Borgialli, "Piccole storie di strada: analisi manuale e informatica di testi narrativi", BA dissertation, Università di Torino, a.a. 2006-7) and partly falls within the repertoire I have been collecting over the years for teaching purposes. Most materials are also available in paper format, but the online version was preferred with a view to electronic querying. The files containing only verbal, first-person narrative (excluding the editorial apparatus, visuals, and reported stories) amount to 40600 words. Concordancing software: Tact.

[6] See for instance Cook Gumperz (2005: 249): "The story is told in response to a question although it looks monologic […] the interviewer's question […] provides the introduction/orientation for the story in terms of the Labovian paradigm".

stance of the teller. What needs to be stressed here is the value of these documents in spite of their truncated nature. Notwithstanding the omission of the conversational event in which they were originally embedded and of the consequent linguistic, cognitive and emotional dimensions of the co-constructed activity of the telling, it is still possible to capture what we are looking for, that is the children’s perspective on their own lifestyle and its story. In order to avoid the emotional entrapment of readers, editorial policy may set aside the voice of the interviewer/co-teller and focus on ‘facticity’, producing an emotionally aseptic sequence of minimal ‘reporting’ sentences; or it may razor down a story to what the adult stance sees as the moral momentum or punch-line in the narrative. For example, the following excerpt unequivocally sets off the so-called ‘Calcutta syndrome’, i.e. the quickness with which the children learnt to pick and choose the organisations that best served their survival needs, without modifying their attachment to the street:

> Organisation B is for eating, organisation C is for sleeping, and organisation D is for playing. (Annapurna, 8 years old)
>
> (www.savethechildren.net/nepal/key_work/street_children.pdf)

Yet, in spite of their being stripped of all interactional features, these accounts do evoke sets of relationships and the teller’s positioning within these relationships. Individual voices, in other words, animate a sense of the self within a specific setting; collectively, these voices constitute multiple versions of the same basic tale of struggling, construing a scenario of heroes and villains, where not only the lack of emotional outpouring, but emotional restraint and even ‘hyper-silencing’ or emotional self-censorship bespeak a social world so dismal that it is implicitly judged by the teller as ‘closed for talking’ (Valsiner 2004). Closer reading of these accounts will reveal their value as counter-narratives, resisting an inscribed archetypal narrative and construing the identity of the victim who liberates him/herself from the dominant narrative of the caring adults and from the myth of the obedient child.

4. Counter narrative one: nurturant figures

A number of the stories in the corpus are told by orphans ("I don't have no mother or father"; "my mother/father died/is dead/has passed away/was shot/were killed") and by abandoned children ("my mother abandoned me/left/ left home/got angry and went/ broke up with my father and disappeared"; "my father is in jail/lost his job/ran out of money/ married another woman/was taking drugs"). The remaining stories have one striking feature in common: the beginning of beginnings is not perceived and told as the identification of one specific place of birth, but as the identification of one or more figures that loom large in the recall of the past for their deviant behaviour. Mothers, fathers and vicarious figures – grandmothers, stepfathers, aunts, uncles, adults (usually an elderly woman) to whom the child was sold or sent to as a house servant – who should have been nurturant figures or who are culturally expected to provide material and emotional nurture, but fail miserably and, further, often drown their sense of failure and inability to cope into any kind of cheap alcohol. The association of degrading poverty with alcohol is the root of most child battering, child abuse, physical and psychological torture: the orientation of the story is thus formulated as agency on the part of the child, i.e. running away, escape, self-liberation, from brutality in the home. In other words, the prototypical story of the street child begins with a flight from the archetypal nurturant figure turned into an archetypal figure of persecution.

However, there is a need to find a justification for this deviant turn. Thus, topics such as poverty and alcohol are in most cases associated with the 'escape story schema' which in Labovian terms coincides with the initial orientation of the story. This detailing of circumstances in connection with violence usually precedes the telling of the violence, revealing a kind of compassion which deserves further comment.

The early experience of violence is subsequently formulated outright through the use of the material process *BEAT/BEAT UP* in the active voice, the child being recipient (*ME*). Compulsive repetition of brutal acts is actualised through time circumstantials ("always", "all the time", "at any time", "every day") or such aspectual devices as *USED TO* and *WOULD* (“used to beat me up”, “he’d beat me”); the gratuitous nature of the violence (“beat me for no reason”, “for nothing”) is also clear in the child’s awareness, but it is the manner of the beating that becomes prominent, detail growing from circumstances of manner (“a lot, like a dog”, “savagely”, “pretty badly”, “too much”, “very badly”, “to death”) to include the actual instruments with which the abuse was perpetrated (“electric cord”, “dried cow hide”, “a stick or board”, “a skillet”, “wire”, “fist”, “pots”, “pans”, “skillet, with whatever they got their hands on”). In the case of the *restavek* – slave children in Haiti, who are sold into bondage to a ‘godmother’ or ‘aunt’ – the telling often includes circumstances of manner and their permanent physical consequences, from scars to burns to injured or amputated limbs, which are pointed to (a visual confirms the veridical nature of logodeixis) or, given their prominence in the child’s recall, are expanded into a whole segment of the story – see for instance the following excerpts from the Haiti Streetkid Gallery:

> Then a lady came and bought me from my uncle and ever since, I've been going through hell. She used to put me on my knees with a big rock on my head and one in each hand. Sometime, she put hot gravel under my knees to make me kneel on.
>
> I have this big scar on my forehead where she would hit me with a hot skillet when she was mad. She did this many times. I have headaches almost everyday now because of this. One time when she was really mad, she dropped a big iron pot full of hot grease from frying chicken onto my hand while it was still in the wash basin. I thought it cut off my hand, but it only crushed it and burned it real bad. She wouldn't take me to the hospital because she might get in trouble. She only wrapped my hand in rags until it got well.

Lexicogrammatical co-selections of the material processes *BEAT* and *HIT*, and manual inspection of the corpus for more detailed descriptions of abuse, reveal that family violence is crucial in each child’s decision to take to the streets. This first and fundamental instance of agency is told in the first person, for, in this corpus at least, fleeing is a personal decision which follows upon the child’s full realisation of being treated as property and as a scapegoat. The formation of a revolting conscience is the mover in the ‘escape’, where ‘flight’ is resolution to the earliest stumbling block: violence and abuse in the native family or early context of socialisation. To be noticed is a diversification of cultural scripts of motherhood where the biological mother or father can be replaced by other nurturant figures, but common to such figures is a deviance generating counter-narratives of the master tale (Andrews 2004a) of the ‘normal’ family.

The ‘resolution’, in terms of Labovian narrative syntax, is a choice for the ‘lesser’ trauma of self-exclusion based on tension with the underlying script of the caring family. And the tension with the master narrative is not necessarily entirely oppositional (see Andrews 2004a: 2), for these children’s perspectives indeed inscribe the ‘facticity’ surrounding their decision in a contextual frame of circumstantial explanation for adult cruelty.

There is an attempt, in other words, not to posit one’s agency in terms of a Manicheian opposition to an idealistic, de-contextualized mother figure. Rather, the narrator ‘I’ is positioned within a moral order where degrading social practices are nearly ‘naturalized’ by absolute poverty. Ignoring such circumstances and openly denigrating the mother with unforgiving righteousness would interfere with the narrative shaping of a self which can achieve continuity (Andrews 2004b: 55 and ff) and grow beyond its roots of shame. That such justification is at the same time self-directed in the effort to resist and contest deep-seated feelings of worthlessness should not be surprising. For these young heroes, having found in themselves the strength to escape, discover that the big world is no less hostile than the social world they left behind, and just as prone to inflict the feelings of being unloved and unwanted. Particularly powerful in this respect is the all-encompassing metaphor used by a young narrator to condense the world’s attitude as perceived by this outcast group: “We’re just like human garbage” (see epigraphs, this study).

5. Counter-narrative two: the lawless police

As the runaway and other categories of children with loose family bonds face the harsh realities of fending for themselves in the urban environment and develop a “special relationship to the street”

(Glauser 1997: 146), with the broad spectrum of variants well described by Glauser, they coalesce with other children in groups whose dynamics can be both supportive and exploitative. At the same time, their presence in public places makes them an unseemly sight to the mainstream desiring 'clean streets', while the activities they practise to cope with basic material needs may at times have thin borders with the illegal. Thus, the ubiquitous cop in the role of anti-hero:

> The police treat us badly. They hit us. Not for any particular reason … just because they feel like it. They've hit me lots of times. They hit with their rifles, or with sticks, on our backs and stomachs. And sometimes they just punch us in the stomach with their hands. They also take our paint thinner and pour it over our heads. They've done that to me five times. It's awful, it hurts really bad. It gets in your eyes and burns; for half an hour you can't see anything. (www.hrw.org/reports/2001/children/5.htm, on 4/10/2007)

The heavy-handed behaviour of police and of the entire judiciary in many countries, particularly since the matter of children's rights in terms of protection by the law and rights to due process in case of law infringements has been regulated by the Convention on the Rights of the Child*,* has been covered by innumerable press reports, by well-known international NGOs and by the UN, a document of paramount importance being the Report to the UN Commission on Human Rights on "Civil and Political Rights, Including the Question of Disappearance and Summary Executions", particularly its addendum entitled "Mission to Honduras" (see Cortese 2005).

Thus the myth of the protective patrolman, which looms large in the child lore of the North, shows its opposite face, that of the corrupt officer, in the South as much as in the transition economies and the other parts of the world where old and new poverty is engulfing children into homeless situations.

The 73 occurrences of police in the present corpus can be arranged into a crescendo of terror, building up from the plain statement "the police came". To the youth using this language, *CAME* extends into a negative semantic prosody[7] involving being

[7] Semantic prosody: "The consistent aura of meaning with which a form is imbued by its collocates" (Louw 1993: 157).

addressed brutally (*CURSE*, *INSULT, ACCUSE*), being physically harassed (*CATCH, GRAB, PICK UP, BUNDLE, TAKE TO THE CAMPS/TO THE STATION*), deprived of liberty (*ARREST*, *PUT IN JAIL*), robbed (*EXTORT*, *STEAL*), rape (*FORCE US TO SLEEP WITH THEM*). These physical processes, mostly in the passive voice, instantiate and amplify a semantic preference for the notion of a persecuting force exercised not only on the individual but more frequently on the group: the narrative in most cases proceeds from the logodeictic centre *US,* typically fronted by *THEM.* But fear is not the only dimension of this negative aura (see footnote 7): there is also the certainty that no maltreatment will be dealt with by the police if the victim is a street child, e.g. a girl child who had been "gang raped" comments that case documents "go missing". The extreme negative prosody is constituted by torture at the hands of the police, from beating to the details exemplified here in the epigraph excerpts. The coming of the police is systematically placed in a sinister light, creating sinister expectations, not only through the co-selections in single concordance lines, but through longer stretches of text. Thus, 'police' becomes a truly loaded word in an essentialized, 'us' *vs* 'them' confrontation.

Only one case in the entire corpus involves a supportive policeman. It is a reported story and has been included in a separate data base – this is worth emphasizing, for none of the narratives of personal experience in the data base used for this study include instances of a policeman inspiring trust:

> Makinzie lived on the streets for several years after the death of his parents. One night while sleeping in an abandoned service station, his face was doused with kerosene and set on fire by some unknown passerby. He was not taken to the hospital for over three days when a high ranking police officer happened to notice him sitting on the side of the street.
>
> (quicksitemaker.com/members/immunenation/StreetkidGallery.html)

The abominable cruelty here lies of course with the action of the "unknown passer by", but the qualifier "high ranking" should not go unnoticed either, since the implication is that no other passer by or lower ranking policeman gave the atrociously hurt human being a second thought.

In short, the collective voice of the street children narratively essentializes the police as the revolting Other, the arch-enemy lurking around the corner to snatch you into the unfathomable world of anguish which 'normal' children displace in fairy tales – for street children the bogeyman is real, night and day.

If, as Hymes (1996: 33) has so clearly stated, the social use of language involves on the part of the analyst the task of understanding its meaning to its local users, the collocations of *POLICE* and the large predominance of passive morphology in colligations lead us to read in the concordancing of this key word a stable fear on the children's unstable, precarious life horizon. Their intersubjective construction of police does not have much in common with the notion of the police construed by Western children, for in many geopolitical areas the question of citizens' safety still applies only to the privileged few.

6. Conclusion

This study set out to compare the literature about street children with the narrative experience of street children, with regard to problematic areas in the life trajectory of this particularly vulnerable group, which mark their life as outcasts and challenge their resourcefulness, agency and coping skills from very early on. These are the family and the police forces as representatives of a society's care for the safety of its members. In both cases, the children's voices are in full consonance with adult professionals involved in the implementation of children's rights. Blatant violations of children's rights occur in the native home – or, where the notion of family extends to near and distant relatives – to the home that is expected to provide physical and emotional nurturance, forcing children into the loneliness and despair that make running away seem the only viable solution. But the promises and attractions of the street scenario, with its new attachments, new bonds and new challenges for survival – the street subculture – is fraught with dangers of violence coming partly from unsuspected quarters, such as older youth in the street gangs, partly from the offensive neglect of the mainstream 'passing by' and hurting if possible, and so very largely from the uniformed yet lawless representatives of the law.

The circumstantial details on adult wrongdoing and brutality deserve attention, not only for the literal content as evidence of human rights violations, but for the salience of the body in these children's recall of the experience of violence. Their personal narratives move from mental and emotional as much as from physical memory of suffering, forever written on their scarred and injured bodies. Their reminiscent bodies enter textual practice, as it were, deploying their expression of stance through a code that is more than verbal.

Though not particularly rich in affective disclosures and hard to analyze in its decontextualized form, the narrative discourse of the children does however construct and construe a social positioning, more ambivalent and complex in the case of the relationship with the native social niche than in the relationship with the representatives of the law, who are steadily and straightforwardly projected as the arch-enemy. But the fracture from the family, the act of self-eradication, often bears a subtext of regret, which explains the toing and froing of many street kids who try to return home, invariably to find that their self-concept has changed and that their memories of the native context no longer match its present shape. No matter how far their outward journey to the streets of the city, they keep carrying an innermost sense of something 'missing', the coveted nurturing love they never received or received only too briefly, which has such a distinctive ring in the sweet and bitter evaluation in the (Labovian) coda to Ricardo's story:

> If I could wave a magic wand and change one thing in the world I would have my mum live with my dad again. My dad is not so good. My dreams for the future are that I could work in a bakery, live in a proper concrete house, have a wife and start my own family. I want 15 or 20 children! But in reality none of this will happen to me. *Nothing that I want can come true.*
>
> (www.newint.org/issue 377/ricardo.htm. My emphasis)

References

Andrews, M. (2004a) "Opening" (pp. 1-6); "Memories of mother. Counter-narratives of early maternal influence" (pp. 7-26), in M. Bamberg and M. Andrews (eds), *Considering Counter-narratives. Narrating, Resisting, Making Sense*, John Benjamins, Amsterdam.

Andrews, M. (2004b) "Response to commentaries on 'Memories of mother. Counter-narratives of early maternal influence'", in M. Bamberg and M. Andrews (eds), *Considering Counter-narratives. Narrating, Resisting, Making Sense*, John Benjamins, Amsterdam, pp. 51-59.

Bourdieu, P. (1991) *Language and Symbolic Power*, Polity Press, Cambridge, Chapter Two, pp. 66-89. (*Ce que parler veut dire. L'économie des échanges linguistiques*, Fayard, Paris, [1982], pp. 59-95).

Boyden, J. (1997) "Childhood and the policy makers: a comparative perspective on the globalization of childhood", in A. James and A. Prout (eds), *Constructing and Reconstructing Childhood*, Routledge/Falmer, London, pp. 190-216.

Cook-Gumperz, J. (2005) "Institutional memories. The narrative retelling of a professional life", in U.M. Quasthoff and T. Becker (eds), *Narrative Interaction*, John Benjamins, Amsterdam, pp. 244-261.

Cortese, G. (1995) "Language and gender in conversational narratives", in E. Siciliani, A. Cecere, V. Intonti and A. Sportelli (eds), *Le trasformazioni del narrare*, Schena, Bari, pp. 23-54.

Cortese, G. (2004) "Pro-social advocacy on the web: the case of street children", in C.N. Candlin and M. Gotti (eds), *Intercultural Aspects of Specialized Communication*, Peter Lang, Bern, pp. 283-309.

Cortese, G. (2005) "On children's right to life: virtuous management of intercultural conflict", in G. Cortese and A. Duszak (eds), *Identity, Community, Discourse. English in Intercultural Settings*, Peter Lang, Bern, pp. 453-484.

Cortese, G. (2007) "The right to be just *other* children: protectionist and liberationist ideologies in the discourse of children's rights", in G. Garzone and S. Sarangi (eds), *Discourse, Ideology and Ethics in Specialized Communication*, Peter Lang, Bern, pp. 73-99.

Glauser, B. (1997) "Street children: deconstructing a construct", in A. James and A. Prout (eds), *Constructing and Reconstructing Childhood*, Routledge/Falmer, London, pp. 145-163.

Hymes, D. (1996) *Ethnography, Linguistics, Narrative Inequality. Toward an Understanding of Voice*, Taylor & Francis, London.

James, A. and A. Prout (1997) "Re-presenting childhood: time and transition in the study of childhood", in A. James and A. Prout (eds), *Constructing and Reconstructing Childhood*, Routledge/Falmer, London, pp. 230-50.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

Labov, W. (1972) *Language in the Inner City*. *Studies in the Black English Vernacular*, University of Pennsylvania Press, Philadelphia.

Labov, W. and J. Waletsky (1967) "Narrative analysis", in J. Helm (ed.), *Essays in the Verbal and Visual Arts*, University of Washington Press, Seattle, pp. 12-44.

Le Roux, J. (1996) "Street children in South Africa. Findings from interviews on the background of street children in Pretoria, South Africa", *Adolescence* 31 (122), pp. 423–431. www.pangaea.org/street_children/africa/safrica2.htm

Louw, B. (1993) "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies", in M. Baker, G. Francis and E. Tognini Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam/Philadelphia, pp. 157-176.

Valsiner, J. (2004) "Talking and acting: making change and doing development", in M. Bamberg and M. Andrews (eds), *Considering Counter-narratives*. *Narrating, Resisting, Making Sense*, John Benjamins, Amsterdam, pp. 245-256.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 71-84 © 2008 Polimetrica International Scientific Publisher Monza/Italy

The illocutionary force of interrogatives in English varieties

Roberta Facchinetti – University of Verona

> Sometimes when someone is asking you a question you can become Socratic, and ask them a question, and have them answer their own question for them.
>
> (Vincent Bugliosi)

1. Introduction

In any communicative event, the message conveys not only content but also a set of subjective aspects pertaining to (a) our attitude to that content and (b) our relationship with our interlocutor(s). This leads to an interlocking of linguistic and non-linguistic features all contributing to the overall illocutionary force of the message. As is well known, such pragmatic force may not necessarily be in a biunivocal relationship with the surface message, since certain speech acts – like requests and orders – are intrinsically face-threatening and thus often require strategic redress.

A typical structure that seems particularly prone to the discrepancy between superficial form and illocutionary force is the interrogative clause; indeed, in all languages questions may actualize some deeper illocutionary aims, ranging from polite offers to orders, complaints, and even rebukes. So, for example, a polar question like "will you have something to eat?", is implicitly interpreted by the interlocutor not as an interrogation, but rather as an offer or invitation. Similarly, by uttering "don't you think it's stupid?", the speaker does not actually ask for information but rather seeks confirmation of his/her subjective point of view.

This subjectivity of intent may be expressed even more overtly by means of a peculiar pattern of interrogatives, with modal verbs

(ModVs) modalising mental verbs (MentVs), as in “will you think that”; indeed, both MentVs and ModVs are generally acknowledged as means of expressing different degrees of subjectivity. It is this specific syntactic cluster, namely ModV + MentV in interrogative contexts, which will be the main focus of the present paper. Specifically, I intend to analyse the central ModVs collocating with the MentVs *KNOW*, *SEE*, *THINK*, and *WANT* – which are also among the most frequent lexical verbs in English (Biber *et al.* 1999).

Undoubtedly over the last few years scholarly research has been thriving on culturally-bound and sociologically-based differences concerning the illocutionary values of interrogatives. Cross-cultural pragmatic specificities have been highlighted, for example, by Blum-Kulka (1987) between English and Hebrew, and by Fukushima (2000) between English and Japanese. In turn, Upadhyay (2003) has focused on the actualization of requestive acts in Nepali, while García (1993) has analysed male and female Peruvian Spanish speakers.

Yet, to my knowledge, little has been studied so far with reference to possible differentiations of interrogatives within different varieties of ‘the same’ language. To bridge this gap at least partially, I have focused on the seven components of the International Corpus of English (ICE) currently available: British English (GB), New Zealand English (NZ), East African English (EA), Hong Kong English (HK), Indian English (IND), Philippine English (PHI), and Singapore English (SIN).

Each corpus covers 1 million words distributed across written and spoken samples from a number of textual varieties, thus totalling 7 million words. Since the corpora are structured exactly in the same way, they are all comparable, which will help me identify possible discrepancies in frequencies and uses of the collocations under scrutiny among the corpora themselves.

By analysing ModV + MentV collocations in these corpora, I intend to verify (a) to what extent the superficial structure of interrogatives may diverge from the actual illocutionary force of the utterance in different varieties of English, (b) what types of illocutionary acts are conveyed, and (c) if such illocutionary acts are evenly represented in the seven ICE corpora taken into consideration or rather if they diverge from one variety to the other.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

2. Distribution of ModV + MentV collocations

As shown in Table 1, the data testify to the fact that only a few central ModVs co-occur freely with the four MentVs under scrutiny; indeed, *SHALL* and *MUST* have not been recorded, while *MIGHT* appears once only in GB:

(1) Might she have seen him first, looking shabby and wretched, and have decided that this was the moment when she no longer wanted to be associated with him? (ICE-GB:W2F-008 #93:1)

The virtual absence of *MIGHT* is understandable since in itself the pattern '*MIGHT* + lexical verb' (independently of the type of clause where it occurs) is particularly rare in GB and is even totally absent in the other six varieties.

Aside from this, however, it is interesting to notice that the collocational patterns under scrutiny favour verbs of possibility and willingness like *CAN* and *WILL*, while they tend to avoid verbs of obligation like *MUST*, as highlighted above, and *SHOULD*. Considering that *KNOW*, *SEE*, *THINK* and *WANT* are themselves among the most frequent lexical verbs in English, the discrepancy in occurrence of the expressions is quite possibly due to the peculiarities of the different lexical bundles in interrogative contexts and also to their semantic-pragmatic functions.

Corpora	Can	Could	Will	Would	May	Should	Total
ICE-EA	4	3	3	7	1	1	19
ICE-GB	16	3	2	13	-	-	34
ICE-HK	28	1	6	17	-	-	52
ICE-IND	8	3	3	8	16	-	38
ICE-NZ	20	2	2	12	-	2	38
ICE-SIN	25	2	3	14	4	3	51
ICE-PHI	8	1	8	14	14	1	46
Total	109	15	27	85	35	7	278

Table 1. ModV + MentV in interrogatives (raw frequencies in the seven ICE corpora).

As illustrated in Table 1, *CAN* and *WOULD* are undoubtedly the modals that occur in harmonic collocation with the four MentVs most frequently. The figures are partly justified by the fact that both modals have been proved to be respectively the second and third most frequent ModVs in English (Biber *et al.* 1999: 486). Strangely enough, however, *WILL*, which Biber *et al.* place at the very top of the frequency scale for ModVs, is among the least frequent when collocating with MentVs, with its 27 occurrences, thus outnumbering only *COULD* (15) and *SHOULD* (7).

Considering the distribution of the data according to varieties, while EA (19 occurrences) and GB (34), seem to behave similarly, HK (52), SIN (51), and PHI (46) exhibit the highest numbers of recordings. Indeed, Asian languages are generally claimed to be more keen on the expressions of face-saving ways of speaking (Hinkel 1997; Closs Traugott and Dasher 2002; Haugh 2007, among others); such culturally-determined feature may have been retained in the local varieties of English as well.

Finally, it goes without saying that the peculiarity of the structure under scrutiny has necessarily affected the context of occurrence of the data, which have all been recorded in interactional extracts, be they the spoken medium or private letters, demonstrations, academic writing[1] or again fictional dialogue.

3. Illocutionary functions of interrogatives

Three different pragmatic values have been recorded: (a) request for information, (b) request for action, and finally (c) rhetorical question signalling emotional involvement. The data from all the seven varieties taken together have yielded the following results:

[1] Though limited in number, due to the restricted type of collocation analysed in my corpora, demonstrations and academic writing in particular appear to be privileged contexts of use for interrogatives, since they create anticipation, arouse the interlocutor’s interest, and bring him/her into some sort of dialogue with the speaker/writer (Webber 1994, Hyland 2002).

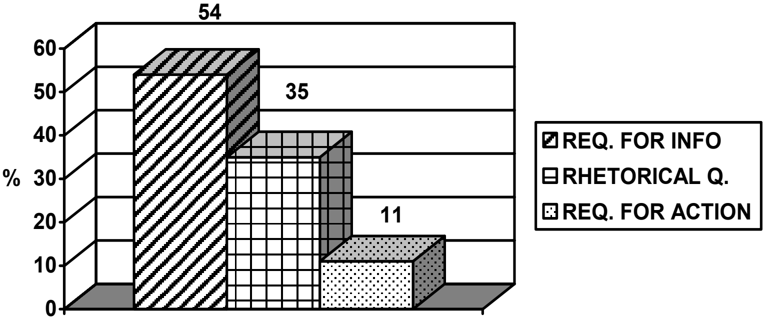


Figure 1. Illocutionary force of ModV + MentV in interrogatives (percentages, all ICE corpora).

While requests for information account for the large majority of the occurrences, requests for action appear to be the least frequent of all. A possible cross-reference with data from languages other than English might provide interesting clues about the pragmatic functions of these interrogatives in terms of cultural differentiations and structural idiosyncrasies.

As will be seen in the sections to follow, each of the three illocutionary values identified in the corpora are strongly bound to a limited set of collocational patterns and cannot be generalised to all varieties. Let us focus on each one of them in detail.

Requests for information

Requests for information are supposed to be the most straightforward type of questions, since there is an apparently direct relationship between the semantic value conveyed by the superficial form and the actual illocutionary force of the utterance. In my corpora, they total 54% of all ModV + MentV collocations and are mostly actualized by "will/would you think/want" and by "can/could you see". Both collocations are particularly frequent in the varieties GB, SIN, HK, and PHI:

(2) When **will you want** the microvawe oven delivered? (ICE-NZ:W2D-011#130:1)

In instances like (2) above, *WILL* seems to be a mere alternative of the auxiliary *DO* and appears to be exploited possibly to convey more formality and to strengthen the 'willingness' feature; in contrast, the tentativeness of *WOULD* reduces the degree of factuality and makes politeness more prominent:

(3) Do you like teaching what age level children **would you want** to teach (ICE-HK S1A-001#X523:1:Z)

A similar situation occurs when the four MentVs collocate with *CAN*, particularly in the pattern "can/could see"; here the modal appears to have an empty value, which, unsurprisingly, accounts for the overwhelming majority of *CAN/COULD* + MentV in the corpora under scrutiny:

(4) <B> And what do you see being sold
<C> I can see a woman over there selling bananas
<B> Yes that woman over there is selling bananas What else can you see
<C> Uh I can see some people taking tea and mandazi in a kiosk
<B> Oh yes You are very observant (ICE-EA:S2B073K)

This pattern is particularly frequent in SIN, HK, GB, and NZ and it can also take on a less overt discursive function; let us consider (5):

(5) **Can you see** the figures are the same there's no change (ICE-SIN:S1B-071#104:1:A)

Here the speaker asks the interlocutor if he is capable of seeing that "the figures are the same", in other words he points to a certain situation and at the same time also indicates the situation itself; in so doing, he solicits the addressee's agreement and leads the way though not imposing his stance overtly. In a study on this ModV + MentV collocation in affirmative contexts (Facchinetti and Adami, forthcoming), this discourse pattern has been identified as typically intersubjective, since the speaker leads the way by asking the interlocutor for confirmation of a state of fact or of a personal conviction the speaker himself/herself has. Demonstrations, speeches and commentaries in all the seven varieties under scrutiny seem to be particularly prone to this rhetorical function of “can see”.

Alongside with “can/could you see” and “will/would you want/think” – which are relatively straightforward since they tend to convey a direct biunivocal relationship between superficial structure and illocutionary force – another frequent, more peculiar collocational pattern is “may I know”, as exemplified in (6):

(6) Sir **may I know** your name please (ICE-IND:S1A-067#31:1:A)

The *Oxford English Dictionary* records this cluster under the value of “expressing permission or sanction: to be allowed (to do something) by authority, law, morality, reason, etc.” (OED online version). Similarly, Huddleston and Pullum (2002: 183) specify that “in questions, subjectivity involves the addressee as deontic source, as in ‘may/can I attend the lectures’, asking for your permission”. Interestingly, in all the occurrences recorded in the ICE corpora “may I/we know” expresses the illocutionary value of asking for permission only at a superficial level, since the actual function is a request to the addressee, who is required not to grant permission, but rather to provide details and information on a topic/issue. Furthermore, in the majority of the cases recorded, “may I know” is exploited as a hedge introducing the real question, possibly to reduce the impact of the infringement:

(7) The question is simple: With all due respect, **may we know** what equipment, machinery, fixtures are serviced (ICE-PHI:W1B-028#77:4)

This expression has been recorded almost exclusively in IND, PHI and SIN, thus further confirming the more formal and polite distinctiveness of some of the Asian varieties, as opposed to GB and NZ, for example, where no instance of it has been recorded.

Requests for action

According to a recent study by Clayman and Heritage (2002), requests structured as “will you” are more frequent than those formatted with “can you”, particularly in press conferences; indeed, appealing to the recipient’s willingness is more deferential and less direct than openly addressing the recipient’s ability. In collocation with MentVs, however, my data have yielded different results, since – independently of the variety – *CAN YOU* + MentV is the most recurrent pattern, while *WILL YOU* + MentV is exploited to a very limited extent.

When expressing a request for action, *CAN* and *COULD* occur almost exclusively with the MentV *THINK*, either with subject *YOU*, as in (8), or with inclusive *WE*, as in (9), which has been recorded to be particularly frequent in EA:

(8) <Z> **Can you think** of an example
<A> No
<Z> Oh come on Think for it
(ICE-HK:S1A-057#X414-417:2:A-Z)

(9) **Could we think** of any medium which they used to communicate with the natives? People like Krapf, Rebmann and Livingstone were here as early as the 19th century, what language did they use to familiarize themselves with the natives and African chiefs alike? (EA:W1A007T)

In contrast, PHI appears to exploit particularly the pattern "may I see", which is similar to "may I know" discussed above. In (10), rather than asking the patient to open his mouth, the doctor asks for permission to see the patient's throat and then takes it for granted that the patient will tacitly agree to do so, by asking him to carry out another action with his mouth open ("say ah"):

(10) **May I see** your throat Say ah (ICE-PHI:S1B-073#146-7:1:A)

Hence, by structuring the interrogation with "may I see", the aggression of the speaker onto the interlocutor is elegantly tempered when asking him/her to carry out an action, in the same way as "may I know" is exploited to mitigate the force of the request for information.

The same politeness strategy is actualized by means of the conventionalised indirectness of "will/would you want":

(11) Now I'll take on some questions Will you want to say a few words in uh Mandarin Uh I think maybe say a few words in Mandarin if you want to (ICE-SIN:S2B-048#61-2:1:A)

Though present in the data under scrutiny, the cluster “will/would you want” is exploited much more sporadically than “can you think” and “may I see”. Unfortunately, since the occurrences of requests for action in all the ICE corpora account for only 11% of all the data, no further generalization can be made on the pragmatic value of such three patterns.

Rhetorical questions signalling emotional involvement

A rhetorical question is generally qualified as a type of interrogation that is not designed for a response, since the speaker is not in search for information but is rather taking a position; as such, it does not presuppose an answer, but signals emotional involvement (Frank 1990; Heritage 2002, 2003; Heinemann 2006, among others).

Interestingly, the vast majority of the questions categorized as ‘rhetorical’ in the ICE corpora appear to have been formatted as negative-interrogatives; this type of clause has been qualified as “biased” (Huddleston and Pullum 2002: 867, 879-880) since it has the complex pragmatic effect of challenging the possible negative outcome of the situation and suggesting the subjectivity of the speaker’s inclination towards a positive answer.

Though limited in number, the data recorded in the ICE-corpora testify to the fact that negative-interrogatives tend to be exploited within the coherence of discourse as a means to confirm the speaker’s attitude. Let us consider the following:

(12) You’re sick, Michael **Can’t you see** that? Are yours the actions of a rational man? (ICE-GB W2F-008 149)

The collocation foregrounds emotional involvement since the speaker emphatically asserts his exhortation to the interlocutor to acknowledge a self-evident truth. Similar data have been gathered by Heinemann (2006) in her study of interrogatives expressing requests; the analysis of her corpus of interactions between Danish home help assistants and their elderly care recipients shows that by uttering negative requests like “can’t I”, “won’t I”, the care recipient appears to feel she is entitled to make such request as something which should be routinely performed.

The cluster “can’t you see” has been recorded in GB, NZ, EA, and, though just once, also in PHI, and in all instances the utterance appears to be particularly forceful, to the point that a flavour of reproach and indignation is frequently present. The same applies when the ModV collocates with *KNOW*, *WANT* and *THINK*, which, however, tend to co-occur more regularly with the tentative forms of modals, thus making the utterance less forceful, as shown in (13)-(14) respectively from ICE-SIN and ICE-NZ:

(13) Has she contacted you yet? I’ll be back in Singapore around 25th of July. My parents come down on the 30th of June and leave on the 20th of July. My convocation is on 16 July & I just hope I do well enough to get my 2-1 - - please, please, please! You know, when I was talking to Fizh, she asked me how Sheeta was & I was kind of surprised & I said **shouldn’t you know** better? And she said Sheeta hasn’t been going home as often as Shaq has, & that Fizh didn’t know what Sheeta was up to - - she was going to contact her that afternoon. (ICE-SIN:W1B-006#92-97:3)

(14) **Wouldn’t you think** they’d just sit down and talk it out?’ Evvie felt a twinge of disloyalty even listening to Polly’s criticisms. (ICE-NZ:W2F-014#58:1)

Although the grammatical subject *YOU* makes the intersubjective force of the utterance more overt, instances have also been recorded of negative-interrogatives with third person subjects, in collocation with *WOULD* and *COULD*, still embodying a similar pragmatic prosody:

(15) I arrived home at 9.30 night having being rained on like nobody’s business. Right now I am considering going to see the chief. What did I do? **Couldn’t they think** of a better punishment. All the same I shall not see him this week. On Friday, I was planning to complain bitterly when we met, but I changed my mind because it was at the wrong time and wrong place. (ICE-EA:W1BSK46)

The cluster indicates that the complementary affirmative formulation (“they could think of a better punishment”) is or would have been somehow expected, thus leading the reader to interpret the utterance as an expression of disappointment.

Rhetorical questions are particularly recurrent not only negative-interrogatives, but also in positively-polarised questions, mostly introduced by the adverbials *HOW* and *WHY*, conveying surprise, disappointment and/or disbelief:

(16) **How can you possibly know** that (ICE-GB:S1A-032 084)

(17) you begin raising an international capital **why should you even be thinking** about doing that (ICE-NZ:S2A-049#17:1:S)

Focusing on the interrogatives introduced by *HOW*, Emmertsen qualifies them as "hostile accountability" questions (Emmertsen 2007: 584) and remarks that though being grammatically interrogative, these formats "provide minimally for the IR's (interviewer) achievement of the formal task of asking questions." (Emmertsen 2007: 584). Indeed, both in (16) and in (17), under the guise of calling for the addressee's account for a problematic position or event, the speaker implies that no such event can be given. Hence, the actual message is "you couldn't know" in (16) and "you shouldn't be thinking" in (17).[2] The questions are actually two negative statements implying something as impossible or inadmissible.

4. Conclusion

The foregoing analysis elicits interesting data which allow us to answer at least preliminarily the questions posed in Section 1, namely which illocutionary values are conveyed by ModV + MentV collocations in interrogatives and if such illocutionary values are expressed in relatively similar ways in different geographical varieties of English.

From the distributional point of view, the data show that EA and GB behave similarly in terms of a low frequency of occurrence,

[2] The cluster "why should you think", exemplified in (17), is also discussed by the OED, which qualifies it as "implying the speaker's inability to conceive any reason or justification for something actual or contemplated, or any ground for believing something to be fact" (OED online version). Indeed, in all such cases the author implies that there is no reason for actualization.

while PHI, SIN and HK are the corpora with the highest number of recordings.

Within this framework, almost half of the 278 interrogatives scrutinized are not real questions from the pragmatic point of view, but rather polite orders/suggestions in disguise or comments on the part of the speaker/writer, calling on the interlocutor's involvement.

Specifically, "will you want" and "can/could you see" are patterns regularly exploited as clear requests for information particularly in the varieties GB, SIN, HK, and PHI; however, at times, "can you see" also takes on a peculiar discursive function with the aim of anticipating or introducing the speaker's attitude or opinion. In so doing, the interlocutor is invited to view the topic foregrounded according to the speaker's/writer's interpretation.

In NZ, GB, EA and, though less prominently, in PHI the same expression is exploited in negatively polarized contexts ("can't you see") to convey disbelief, amazement and emotional involvement on the part of the speaker/writer. In other corpora the pattern is not used, thus leading us to the tentative conclusion that these varieties exploit more straightforward and direct ways of expressing one's attitude than others, like IND and SIN. Indeed, these two varieties share different pragmatic patterns, largely due to the cluster "may I/we know", when asking for information in a polite/formal way.

Undoubtedly, this is only a preliminary study and much more must be done particularly with reference to the different pragmatic actualizations of this syntactic structure in the varieties of English. However, right from this preliminary analysis we have found confirmation of two main interrelated facts. In the first place, the four elements (a) ModV, (b) MentV, (c) negative-/positive-interrogative form, and (d) subject type are bound in a force-dynamic relationship which conveys a higher level of potency than the one profiled by each single component of the cluster considered separately. Secondly, as has been illustrated, the culture-specific background plays a key role in the linguistic actualization of the illocutionary values scrutinized and of their distribution across varieties. Both aspects, mostly the second one, definitely need a more attentive screening in further studies, since they compel us to go beyond the limits of viewing a language, specifically English, as a solid monolith, and oblige us to see it as a set of correlated, but independent varieties, each one with its own individuality, autonomy, and culture-specificity.

References

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English*, Longman, London.

Blum-Kulka, S. (1987) "Indirectness and politeness in requests: same or different?", *Journal of Pragmatics* 11, pp. 131-146.

Clayman, S.E. and J. Heritage (2002) "Questioning presidents: journalistic deference and adversarialness in the press conferences of U.S. Presidents Eisenhower and Reagan", *Journal of Communication* 52 (4), pp. 749-775.

Closs Traugott, E. and R.B. Dasher (2002) *Regularity in Semantic Change*, Cambridge University Press, Cambridge.

Emmertsen, S. (2007) "Interviewers' challenging questions in British debate interviews", *Journal of Pragmatics* 39, pp. 570-591.

Facchinetti, R. and E. Adami (forthcoming) "Intersubjective patterns of English modalised mental state verbs", *English Text Construction* 1.2.

Frank, J. (1990) "You call that a rhetorical question? Forms and functions of rhetorical questions in conversation", *Journal of Pragmatics* 14, pp. 723-738.

Fukushima, S. (2000) *Requests and Culture. Politeness in British English and Japanese*, Peter Lang, Bern.

García, C. (1993) "Making a request and responding to it: a case study of Peruvian Spanish speakers", *Journal of Pragmatics* 19, pp. 127-152.

Haugh, M. (2007) "The co-constitution of politeness implicature in conversation", *Journal of Pragmatics* 39, pp. 84-110.

Heinemann, T. (2006) "Will you or can't you?": displaying entitlement in interrogative requests", *Journal of Pragmatics* 38, pp. 1081-1104.

Heritage, J. (2002) "The limits of questioning: negative interrogatives and hostile question content", *Journal of Pragmatics* 34, pp. 1427-1446.

Heritage, J. (2003) "Designing questions and setting agendas in the news interview", in P.J. Glenn, C.D. LeBaron and J. Mandelbaum (eds), *Studies in Language and Social Interaction. In Honor of Robert Hopper*, Lawrence Erlbaum Associates, Mahwah, New Jersey, pp. 57-90.

Hinkel, E. (1997) "Indirectness in L1 and L2 academic writing", *Journal of Pragmatics* 27, pp. 361-386.

Huddleston, R. and G. Pullum (2002) *The Cambridge Grammar of the English Language*, Cambridge University Press, Cambridge.

Hyland, K. (2002) "What do they mean? Questions in academic writing", *Text* 22, pp. 529-557.

The Oxford English Dictionary Online, Oxford University Press, Oxford. http://dictionary.oed.com/

Upadhyay, S.R. (2003) "Nepali requestive acts: linguistic indirectness and politeness reconsidered", *Journal of Pragmatics* 35, pp. 1651-1677.

Webber, P. (1994) "The functions of questions in different medical journal genres", *English for Specific Purposes* 13 (3), pp. 257-268.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 85-98 © 2008 Polimetrica International Scientific Publisher Monza/Italy

The perception of citizenship in the English press

John Morley – University of Siena
and Alan Partington – University of Bologna

1. Introduction

This paper reports research conducted by the two authors for the project IntUne (Integrated and United: A Quest for Citizenship in an ever closer Europe). The project is coordinated by the University of Siena and will last four years starting from September 2005. There are 29 European Institutions and over 100 scholars across Central and Western Europe involved. The project is multidisciplinary in nature and calls on scholars and practitioners from different fields of study: political science, sociology, public policy, media, linguistics and socio-psychology.

The aim of the project is to study the changes in the scope, nature and characteristics of citizenship which result from the political and geographical enlargement of the European Union. The project focuses on how processes of integration, at both national and European levels, affect the three major dimensions of citizenship: identity, representation and the scope of governance. Problems of citizenship are addressed under this threefold approach by looking at the relationship between the general public and élite groups – politicians and senior civil servants – and between the European and the domestic dimensions of political life.

The Media Working Group, of which the authors are members, is composed of linguists from the Universities of Siena and Bologna in Italy, Cardiff and Brighton in Great Britain, the University of Lorient (South Brittany) in France, and Łodz in Poland. Its main scientific activities include, firstly, the collection and analysis for training purposes of a small Pilot Corpus of newspaper and television texts from the four countries involved. After this, and more importantly, the team will collect, mark up and analyse two larger corpora, known as the Main Corpora, of television and newspaper texts from the four countries using the experience gained from work on the Pilot Corpus. In order to ensure the integration of their work into the activities of the other three Working Groups, the Main Corpora are being collected at the same time as these groups are conducting Europe-wide surveys concerning attitudes to citizenship among élite groups and ordinary citizens. It is hoped that there will be connections between the preoccupations of the élite groups and ordinary citizens as revealed by the surveys and the key arguments found in the media. The work which follows here focuses on the particular topic of the perception of citizenship in the English press.

Our methodological approach, Corpus-Assisted Discourse Studies (CADS, Partington 2004) combines both statistical or quantitative analysis with the closer qualitative reading typical of discourse analysis.

2. The corpora

Two sub-corpora were used in this study: one of 400,120 words from the *Guardian*, the other of 122,066 words from the *Sun*. The two newspapers are fundamentally different: the *Guardian* is a quality paper of the liberal-left, whereas the *Sun* is a working-class popular paper with rightwing tendencies. The articles collected were all the news stories, features, editorials and op-eds which were printed on the six days between Monday 21 November and Saturday 26 November 2005. These two sub-corpora were part of the Pilot Corpus mentioned above and this article is a version of a presentation made at a training workshop in Bertinoro (University of Bologna) in February 2006.

3. Preliminary considerations about the corpus

During this week no problems arose which related directly to the European Union. This could be considered a disadvantage in that the English press would be unlikely to devote much attention to EU matters. On the other hand, we might hope to gain some insight into what preoccupies the English press in an ‘ordinary week’. It could certainly be argued that, if a survey concerning how the country was being governed had been taken in the week beginning 28 November 2005, we might have expected the replies to carry traces of what the papers were saying in the week when we collected our data.

4. The perception of citizenship

In order to find out about British citizens’ perception of citizenship it might seem simplest to look up the two words *PERCEPTION* and *CITIZENSHIP* in our corpora. Unfortunately, the word *PERCEPTION* appears in neither of them. While we might instinctively feel that this word is not normally part of the *Sun*’s lexis, we would expect to find it in the *Guardian*. It is not particularly rare; nor is it a word which is not usually found in the language of quality newspapers. If we look at all the words written in the three major quality papers in a year[1] – there are about 100 million of them – then we find that the word *PERCEPTION* occurs 1334 times, or about once a day on average in each paper. It is simply a matter of chance that the word does not appear in the *Guardian* in our week. The word *CITIZENSHIP* – a rather less common word in the quality press occurring 850 times, or about once every two or three days – is found 11 times in the *Guardian*, though it is absent from the *Sun*. Below are the concordance lines of *CITIZENSHIP* from the *Guardian* corpus.

1 where he obtained Austrian **citizenship.** And that was
2 influence, gave up Canadian **citizenship** for a British
3 whole population into common **citizenship.** However, the
4 A true citizenship is a critical **citizenship.**
5 of energetic, critically engaged **citizenship** Ramadan calls
6 monarchy. We use nationalism not **citizenship** to generate a sense
7 culture - a weak tradition of **citizenship.** In place of a
8 republican tradition of strong **citizenship,** he is remarkably

[1] The corpus used is called Papers93 and contains all the words in *The Times*, *Telegraph* and *Guardian* and their sister Sunday papers during 1993.

9 the passport but no substantive **citizenship."** In terms of

10 of what we need today. A true **citizenship** is a critical

11 Communist who once renounced US **citizenship** and unsuccessfully

From this minimal context it seems as though the paper is engaged in a lively philosophical debate about citizenship: we find expressions like “critical citizenship”, “true citizenship”, “a weak tradition of citizenship”, “no substantive citizenship” etc. The problem here is that the lively philosophical debate is mainly to be found (seven instances) in two op-ed articles written on 21 November, which analyse issues of Muslims and citizenship. Below we will discuss similar problems in the *Sun*.

Although the *Sun* makes no mention of citizenship, it does talk of citizens. This is very much in line with what media scholars have frequently observed of the popular press,[2] namely that they personalize news: so the *Sun* does not talk of abstract ‘citizenship’ but of concrete ‘citizens’. It does not, however, talk very much about citizens. There are four uses of the words, once in the title “Citizens Advice Bureau”, once talking about “American citizens”, once discussing “citizens of the world” and once referring to “British citizens” in the context of the theme of illegal immigrants.

The *Guardian* too speaks of citizens: the word occurs 26 times. Some of the words which co-occur with *CITIZENS* – its collocates – are: *THEIR* (5), *ITS* (4), *POLITICAL* (4), *BRITISH* (2), *CRITICAL* (2), *FREEDOM* (2), *RIOTS* (2), *WHITE* (2). Two examples in context are:

(1) In a visit to Beijing that veered between the solemn and the slapstick, President George Bush worshipped at a state-run … church yesterday, in an appeal for China to grant greater religious and **political** freedom to its **citizens**.(news\G21novs46)

(2) Isfahan is not just the increasingly notorious location of a nuclear processing plant; it's also a beautiful city where many **critical citizens** live.(op-ed\G24nov54)

The phrase “perception of citizenship”, by the way, does not appear at all even in the hundred million word corpus, nor are the two

[2] See Sparks (1992); Morley (1998, 2003).

words found meaningfully close together, that is coming within a span of ten words of each other.

These facts lead us to reflect that it is not always the direct question which will get us the answer we need. If we wish to find out from a newspaper about the perception of citizenship, we might be better advised to discover what the particular preoccupations of British citizens are concerning the way their country is being run. Fortunately, this is precisely what newspapers talk about most of the time.

5. Europe and European

Once again, the most straightforward way to discover media attitudes towards Europe should be to look up the two key words *EUROPEAN* and *EU*. We have rather more success with these words. *EUROPEAN* is used 118 times by the *Guardian* and 14 times by the *Sun*, and *EU* 91 times by the *Guardian* and 32 times by the *Sun*.

Looking at *EUROPEAN* first, we find that many of its collocates in the *Guardian* fall into definable semantic sets. The first one is 'European Union institutions': *UNION* (11), *COMMISSION* (7), *AUTHORITY* (4), *CENTRAL* (4), *COUNTERPARTS* (4), *OFFICIALS* (3), *REPORT* (3), *RIGHTS* (3), *STATES* (3); the second semantic set is 'money': *BUSINESS* (5), *SALE* (4), *£* (3); the third might be labelled 'consumables': *COCAINE* (4), *DRINKS* (3), *FOOD* (3). We can further contextualize one of these collocates, the most common one, *UNION*.

1 However, today India produces more mathematicians than the whole of the **European Union.** (features\g22nov24)

2 With a half an eye on the **European Union**, Turkish leaders say they are determined to shed light on an affair many see as an attempt by anti-democratic forces to destabilise Turkey's progress to EU member (news\g26nov)

3 His arrest has led to expressions of concern from the United States and the **European Union**, as well as causing the worst street violence Uganda has seen in decades. (news\g21nov47)

4 investors welcomed the **European Union's** reduction in sugar subsidies as far less severe than initially feared. (news\g26nov)

5 Europe, mon Foyer (Europe, My Home), 100,000 copies of which have been distributed to primary-school children in Belgium. Its aim is to tell children about how marvellous the **European Union** is. (features\g23nov06)

6 Both the **European Union** and US have made offers to reduce support for their farmers but poor countries are not satisfied.(news\g25nov)

7 What happens now? We need Blair to back the **European Union** delegation going to Montreal with a strong and clear statement about the importance of legally-binding targets,(features\g2dd29)

8 the failure of nation states in the **European Union** to respond to demands for open and fair competition," (news\g24nov50)

9 Zimbabwe has poor relations with Britain, the **European Union** and the United States, (news\g21nov45)

10 Later, Houpette takes Lea and Thomas to a sports club to study its regulations. "Not long ago the **European Union** was given regulations such as these," Houpette tells them. "With this new constitution everything will go like clockwork, just like in your club." He explains t (features\g23nov06)

11 Additional capacity will be lost in coming years as generators restrict production and then close down some coal-fired capacity, rather than make the investment needed to meet **European Union** emission rules. (news\g21nov04)

In most of these examples, affairs relating to the European Union are narrated with journalistic objectivity even when describing problems created by agricultural subsidies (example 6), a notorious sore point for the British. Lines 5 and 10 are a couple of exceptions, where a note of sarcasm seems to be present. The stance towards the EU is sometimes favourable, as in 11 where it is portrayed as a body which sets positive standards, and in 4 and 6 where it is at least moving in the right direction, and sometimes critical as in 8 and 1, where it is castigated, respectively for stifling the free-market and for having low educational standards in failing to produce a decent number of mathematicians.

The *Sun* makes less frequent reference to things European, even bearing in mind that the corpus is smaller than the *Guardian* corpus: 0.11 references per thousand words as opposed to 0.29 ptw in the *Guardian*. Six of the 11 references are to European football, two are to the European Commission, one to the European Constitution, one to Eastern European gangsters, one to building nuclear power plants in Finland and one to the European Monitoring Centre for Drugs – we saw that the *Guardian* too was preoccupied with cocaine use. One of the references to the European Commission describes it as "bloated and arrogant". We see from the context – "Not a single Pound must go before the Common Agricultural Policy is scrapped. If that upsets **the bloated and arrogant** European Commission, too bad" – that the modifiers *BLOATED* and *ARROGANT* are placed in a syntactic position which makes them 'given information'. The *Sun* takes for granted that its readers will agree with this description of the European Commission. The *Sun*'s antipathy for the EU is well known in the UK (Coffin and O'Halloran 2006). We can see this clearly in its treatment of the word *EU*, which it uses slightly more frequently than the *Guardian* – 0.26 ptw as against 0.23 ptw. We find a reference to "the faceless godfathers of the EU". We are told that "What the EU cannot accomplish through the front door, it will bring in through the backdoor". We hear of the "EU gravy train". We are told that the "EU is being strangled to death by regulation". There is talk of sacrificing principle to "curry favour with our EU partners". Chancellor Merkel is described as "stirring the putrid EU pot". And, most damningly, the *Sun* thunders:

(3) It is a luxurious home for failed politicians and self-regarding civil servants who want to get their greedy snouts into a bottomless trough. The EU, with its sprawling membership, bureaucracies, regulations and colossal expense, brings no benefit to us that we could not get elsewhere or achieve for ourselves.

It is important to add, however, that most of the vitriol is poured out in one op-ed article by Fergus Shanahan arguing strongly against what the *Sun* was referring to as "surrendering" Britain's EU rebate. In doing quantitative studies, statistics can sometimes be skewed in this way by a small number of texts. The 124 thousand words in our single-week's editions of the *Sun* is on the small side for a modern-day corpus. But we are at the pilot project stage and the Main Corpora will be more than ten times as large and will include texts from the two months before and the month including the surveys. Another point is that the way of working implicit in the CADS methodology adopted by our group – shunting from the quantitive to the qualitative – should alert us to the danger of using raw statistics which are distorted by having been determined by an idiosyncratic portion of the corpus.

6. Problems of governance for the writers of the *Guardian* and the *Sun*

What, then, are the aspects of governance which exercise the journalists working for the two newspapers in this period? If we can identify these, we can presume that the readers of the papers will at least be aware of them in the following week and would probably make reference to them in a survey. We can start from a simple quantitative analysis and look at the WordList which our software, *WordSmith* (Scott 1998), produces from the corpora and puts in order of frequency. After this we need to examine the lists carefully for words which seem relevant to our interests. As linguists we know that the vast majority of the most frequent words will be function (grammatical) words. For instance the first non-function word in the *Guardian* WordList comes at number 55, *PEOPLE*, and there are only five in the first hundred most common words, *PEOPLE*, *YEARS*, *YEAR*, *GOVERNMENT* and *WORK*. Function words are often extremely interesting if we are trying to define discourse types in linguistic terms (see especially Biber 1988), but in this study we are much more interested in non-function, or lexical, words which indicate the topics of the various articles. However, simply looking at wordlists can be deceptive. There are 13,295 types, or different words, in the *Sun* corpus. The 101st – with a frequency of 137, a very common word therefore – is *SHARON*. Does this indicate that the *Sun* is passionately interested in Israeli affairs? When we check the context, an essential part of CADS procedure, we find that only three of the instances refer to Ariel Sharon, almost all the others are references to a female police constable, Sharon Beshenivsky, who was shot dead in an armed robbery in Bradford.

With these caveats in mind, we edited *WordSmith Tools* word lists to produce the forty most frequent relevant words for our current topic-analysis in the *Guardian* and the *Sun*. These we will call our 'topic words'. Firstly, the *Guardian*:

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

N°	Word	Frequency	N°	Word	Frequency
1	goverment	424	21	power	193
2	work	409	22	country	190
3	world	332	23	health	187
4	UK	292	24	city	184
5	London	291	25	money	181
6	children	283	26	white	181
7	public	282	27	national	180
8	British	270	28	office	178
9	police	267	29	market	172
10	home	256	30	state	168
11	need	251	31	case	167
12	report	239	32	president	165
13	company	238	33	local	162
14	political	221	34	business	161
15	women	220	35	war	160
16	Britain	216	36	Blair	159
17	young	208	37	house	159
18	group	205	38	school	159
19	minister	197	39	nuclear	155
20	party	194	40	problem	153

Table 1. Topic words in the *Guardian*.

The first thing that strikes us is that there is quite a large group of words from the semantic field of politics: *GOVERNMENT*, *MINISTER*, *POLITICAL*, *PRESIDENT*, *BLAIR*. The words *HOUSE*, *WHITE*, *PARTY* and *POWER* might also refer to politics but we will need to examine them in context. The presence of the collocates *WHITE* (18), *LORDS* (13), *COMMONS* (12) and *GOVERNMENT* (4) tells us that *HOUSE* is used frequently in the political sense, but it is also often used simply to mean the place where we live. *WHITE*, apart from collocating with *HOUSE*, also collocates with *PAPER*.[3] *PARTY* is employed almost exclusively in its political sense, with the main collocates *LABOUR* (25), *LIKUD* (17), *LEADER* (12), *RULING* (11),

[3] A White Paper is a post-draft stage document proposed for Parliamentary attention.

POLITICAL (9), *CENTRIST* (7), *SUPPORT* (6), *COMMUNIST* (5), *LEFTWING* (5), *RIGHTWING* (5), *ELECTIONS* (4) etc. Just occasionally we get a reference to a festive occasion such as "the annual office Christmas party". The word *POWER*, as we mentioned, refers predominantly to the production of energy, mainly of the nuclear kind: *NUCLEAR* (42), *STATION* (16), *GENERATION* (6), *ENERGY* (5) etc., so has little directly to do with politics. It turns out, however, that the major collocate of *HOME* is *OFFICE*, more precisely *HOME OFFICE*, emphasizing further the *Guardian's* interest in politics.

The next lexical set is that containing words connected with the economy. Typically, the abstract word *ECONOMY* is not included in the top forty – it crosses the line in 947th place with just 48 instances. But we find *COMPANY*, *MONEY*, *MARKET*, *WORK*, *CITY* and *BUSINESS* – examples not of personalization but of reification, characteristic of both the popular and quality press, which in the time-honoured Anglo-American tradition never use an abstract word when a concrete example can be given. We need to check on the collocates of *COMPANY* to verify that the paper is not referring to lonely-hearts clubs. In fact, we find that it collocates with *SHARES* (9), *ENERGY* (8), *ANNOUNCED* (7), *PROFITS* (7), *PRIVATE* (7), *CREDIT* (4), *FLOATED* (4) etc. Sixty-six of the 84 occurrences of *CITY* have an upper-case initial and refer to the financial centre of London. It is also clear that *MARKET* refers to the abstract money market – and not to a place for the open-air selling of cheap goods – from the presence of collocates like *HOUSING* (11), *ADVERTISING* (3), *BULL* (3), *CURRENT* (3), *WHOLESALE* (3) before the search word – or node – and *FORCES* (6), *SHARE* (6), *VALUE* (4), *CAPITALIZATION* (3), *CONDITIONS* (3), *SPECULATION* (2) after the node.

So far we are seeing, rather unsurprisingly, that a major quality paper occupies itself primarily with matters political and economical.

There is then the minor semantic field related to more defenceless members of society: *WOMEN*, *YOUNG* and *CHILDREN*. The words are interrelated as *YOUNG* has the main collocates *PEOPLE* (523), *CHILDREN* (15) and *WOMEN* (11).

The final grouping is that of *POLICE* and *CASE*. *CASE* is obviously a very general noun but the presence of the collocates *COURT* (7), *RAPE* (7), *PROSECUTION* (5), *TRIAL* (5), *CONSENT* (4), *DRUNKEN* (4), *JUSTICE* (4), *EVIDENCE* (4), *JUDGE* (3) etc. makes it clear that many of the uses of *CASE* are concerned with law and order. We will come back to this point when we analyze the *Sun*.

N°	Word	Frequency	N°	Word	Frequency
1	police	199	21	united	54
2	home	163	22	drinking	53
3	Britain	91	23	parents	53
4	children	78	24	money	52
5	kids	77	25	spent	52
6	family	74	26	dad	51
7	hospital	74	27	pounds	49
8	UK	68	28	government	47
9	murder	66	29	health	47
10	husband	61	30	Bradford	45
11	school	61	31	country	45
12	Blair	59	32	body	44
13	court	59	33	son	44
14	death	59	34	chief	43
15	drug	59	35	drugs	43
16	officers	56	36	public	42
17	British	55	37	armed	41
18	child	55	38	case	41
19	cops	55	39	dead	41
20	daughter	54	40	shot	41

Table 2. Topic words in the *Sun*.

The best represented semantic field in the *Sun* is that of law and criminality: *POLICE*, *MURDER*, *COURT*, *DRUG*, *COPS*, *DRUGS*, *ARMED*, *DEAD*, *DEATH*, *OFFICERS*, *BODY*, *CHIEF*, *CASE*, *SHOT*. All the references but two to *OFFICERS* are to police officers – the word *POLICE* is found immediately to the left of *OFFICERS* 18 times – and on the other two occasions it refers to prison officers and immigration officers. *CASE* is a general noun, but in the *Sun* corpus, when the nine examples of “in the case of” / “in case” have been removed, 27 of the 32 remaining instances refer to court cases. Fifteen of the 43 instances of *CHIEF* refer to police chiefs. The week under examination was the one following the fatal shooting of the

WPC Sharon Beshenivsky in Bradford – this explains the 45 references to Bradford. It could be argued that tragic stories of this kind will skew our statistics, especially in such a small corpus. This is clearly true, but it is also evident that the preoccupations of ordinary citizens are determined by this skewing. Nor is it wrong that this should happen. If a young policewoman is shot in the course of her duty, then the problems of law and order will leap to the head of the list of things that worry the *Sun* readers about living in Britain and will be one of the problems that they will want the government to do something about. This would be reflected in any survey made on governance conducted around this time.

A large number of words indicate family relationships: *HOME*, *CHILDREN*, *KIDS*, *FAMILY*, *HUSBAND*, *CHILD*, *PARENTS*, *DAUGHTER*, *SON*. A closer examination of the stories in which these words appear does not, however, present us with journalism which debates family problems in a theoretical way. (An exception might be *CHILD*, which collocates with *PORN* (6), *ABUSE* (3), *CRUELTY* (1) and is found in articles about these problems). As we can see, this kind of approach to the family is quite rare. Rather, the subjects of stories in the *Sun* are seen as family members. This is part of the personalizing approach of the popular press: we are all, or have been, members of a family, so this is our lowest possible common denominator.

We reach a similar conclusion about the very common words *HOSPITAL* and *SCHOOL*. The *Sun* is not engaged in campaigning about the health or education service; it is talking about people in hospital, notably the footballer George Best who gets nine mentions, and children in school, though there are three references to school *BULLIES*, a growing problem. Once again we might argue that this is typical of popular journalism. As Meinhoff and Richardson (1994) note, when discussing the treatment of unemployment in the press, a popular paper is more likely to produce the headline "Jobless Paul in suicide horror" (*Daily Mirror* 30.4.1991) than "Ministers urged to compel jobless to work for benefits" (*Independent* 29.4.1991).

The doings of government are well covered, as the presence of the word *GOVERNMENT* itself and the name *BLAIR* attest, though the *Guardian* devotes proportionally about four times as much space to politics as the *Sun*.

7. Conclusion

We have seen which topics dominate the pages of the English press in what might be considered a normal week in terms of news. The left-wing quality paper the *Guardian* concentrates principally on political and economic matters, while also devoting space to following up the chief spot news of the previous week, the murder of WPC Sharon Beshenivsky. The working class populist paper the *Sun* devotes a great deal of space to the same news story. It would have been interesting to see if a survey carried out in the following week would have indicated law and order as among the most pressing of the concerns of the papers' readers.

A number of media observers have commented on how so-called quality papers like the *Guardian* are coming increasingly to resemble, even to ape, their tabloid rivals (McNair 1999: 46). Be this as it may, one conclusion we can tentatively reach from the admittedly small amount of lexical data analysed here is that significant differences remain between the two newspaper types, if not always in what is talked about but how it is talked about and the amount of attention allocated, including to citizenship and perceptions of citizenship.

References

Biber, D. (1988) *Variation across Speech and Writing*, Cambridge University Press, Cambridge.

Coffin, C. and K. O'Halloran (2006) "The role of appraisal and corpora in detecting covert evaluation", *Functions of Language* 13 (1), pp.77-110.

Diller, H. (2000) "Kenneth Starr and us. The Internet and the vanishing journalist", in F. Ungerer (ed.), *English Media Texts Past and Present*, John Benjamins, Amsterdam/Philadelphia, pp. 197-213.

Meinhoff, U. and K. Richardson (1994) *Text, Discourse and Context: Representations of Poverty in Britain*, Longman, London.

McNair, N. (1999) *News and Journalism in the UK*, Routledge, London.

Morley, J. (1998) *Truth to Tell: Form and Function in Headlines*, Cooperativa Libraria Universitaria Editrice Bologna, Bologna.

Morley, J. (2003) "Reporting politics: hard news, soft news, good news, bad news", in C. Nocera, E. Persico and R. Portale (eds), *Rites of Passage: Rational /*

Irrational Natural / Supernational Local / Global, Rubbettino, Catanzaro, pp. 404-414.

Murphy, A. and J. Morley (2006) “The peroration revisited”, in M. Gotti and V.K. Bhatia (eds), *Explorations in Specialized Genres*, Peter Lang, Bern, pp. 201-215.

Partington, A. (2004) “Corpora and discourse: a most congruous beast”, in A. Partington, J. Morley and L. Haarman (eds), *Corpora and Discourse*, Peter Lang, Bern, pp. 11-20.

Scott, M. (1998) *WordSmith Tools Manual*, Oxford University Press, Oxford.

Sparks, C. (1992) “Popular journalism: theories and practice”, in P. Dahlgren and C. Sparks (eds), *Journalism and Popular Culture*, Sage, London, pp. 24-44.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 99-118 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Job ads and the construction of identity in contemporary English primary education

Martin Solly – University of Florence

1. Introduction

This paper briefly describes how primary school teachers are recruited in England, before focusing on a small corpus of print ads advertising primary school teaching posts. It then examines some of the linguistic realisations and generic patterning of the texts, with the aim of shedding light on the way identity is represented and shaped in/by the ads.

The critical analysis is grounded in the perspectives of socially situated discourse analysis (Fairclough 1992; Chouliaraki and Fairclough 1999; Fairclough, Cortese and Ardizzone 2007), genre-based theory (Swales 1990, 2004; Bhatia 1993, 2004, 2005), specialized discourse (Gotti 2003, Hyland 2004) and studies on the discourse of advertising, with particular reference to print ads (Cook 1992; Lombardo 2001; Solly 2002, 2007).

Primary schools in England

Since education in the United Kingdom is a devolved area and there are separate education systems in England, Scotland, Wales and Northern Ireland, this paper will focus on the situation in England, where compulsory education (for all children aged between five and sixteen) is currently divided into four Key Stages (KS).[1] Primary

[1] In January 2007 the UK Government presented its proposal to introduce the raising of the age until which students must continue to receive some form of education or training to 18 by 2013. An earlier stage, the Foundation Stage, is for

schools are normally responsible for the first two stages, also known as lower primary or infants (KS 1: ages five to seven, Years 1 and 2) and upper primary or juniors (KS 2: ages seven to eleven, Years 3, 4, 5 and 6). The provision of education in England can be private in independent schools or publicly funded in state schools. Here we only look at the publicly funded sector, which in April 2007 at national level was the responsibility of the Government Department for Education and Skills,[2] and at local level of the Local Education Authorities.

The recruitment process

Primary school teachers are usually recruited through job advertisements placed in the national press, in particular in the *Times Educational Supplement*. It needs to be remembered that public sector teaching posts are advertised to the public at large and that the whole recruitment process needs to be carried through. Thus, the teaching post must be advertised and applications received within a set deadline. After applications are received, a shortlist is drawn up of those who will be called for interview (usually up to about six). The shortlisted candidates are then interviewed and the post filled. If no candidate is considered suitable for the particular requirements of the job, it will almost certainly be readvertised. Some local authorities (LAs) operate a 'pool' system for recruitment, in which applications from newly qualified teachers (NQTs) are dealt with centrally rather than by schools advertising vacancies individually.

The Times Educational Supplement

The *Times Educational Supplement* (TES) is a weekly UK publication covering the world of primary, secondary and further education, as well as teaching job vacancies. Published on Fridays it also has an online presence, particularly in connection with its

children aged three to five.

[2] In June 2007 its responsibilities were divided between the new Department for Children, Schools and Families (DfCSF) and the new Department for Innovation, Universities and Skills (DIUS).

employment vacancies advertising. The Jobs section on the website (www.tes.co.uk) is home to all the vacancies listed in the TES paper and is updated daily.[3] According to the company profile available on the website (accessed on 15 September 2007), the TES had an average net circulation per issue for the second half of 2006 of 62,258, with 98.2% of these being sold in the United Kingdom and the Republic of Ireland and an average monthly readership of around 474,000. The readership tends to be concentrated within the teaching profession/educational community: many subscriptions are institutional and a copy is usually available in staff rooms in most English schools.

2. Materials and method

The sample

The sample consists of a small corpus of twenty-two job ads taken from the 29 April 2007 issue of the TES. This date was chosen as being at the height of the annual spring recruitment surge in England, here colourfully described by Roger Pope:

> Once again the recruiting drums are sounding across the land. As if driven by a springtime biological urge, the jobs sections of the TES grow fatter by the week so that by the end of term they have to be taken to the staff room by forklift truck. (Pope 2007)

The language chosen by Pope is resonant in imagery from different spheres: military, “recruiting drums sounding”; natural science, “springtime biological urge”; agricultural, “grow fatter by the week”; industrial, “forklift truck”. In any case he gets across well the idea of large numbers of schools advertising at a specific time of year for large numbers of teachers in bulky newspaper issues; indeed the issue looked at here consists of four large inserts, packed with job ads. Those in the sample are all taken from the *circa* 400 found on pages 102-150, under the heading Primary Education, and

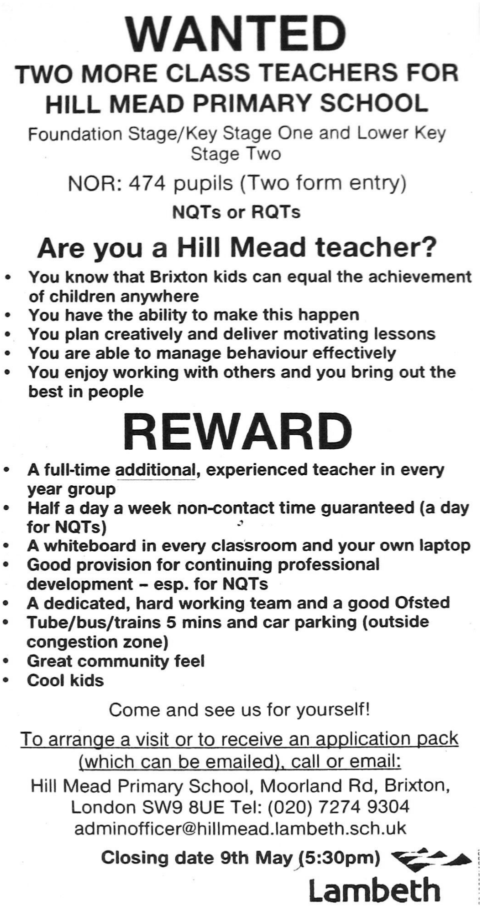
[3] Although the online TES job ads would indeed be a rich source of data for further corpus linguistics analysis, the sample dealt with in this paper only refers to the print version of the ads.

only refer to public sector primary schools. The job ads usually present the post advertised with a brief outline of the skills and experience required. Some are quite short and terse, whilst others provide a considerable amount of information about the school as well as setting out the profile of the prospective candidate.[4] The language in which the ads are couched can be highly specialised and sometimes impenetrable. Wignall, with reference to job ads in the world of business, describes the 'despair' of job hunters as they are plunged into deeper despair, "via the medium of job ads no normal person can understand, written in some kind of 'recruitmentese'" (Wignall 2007).

Method

Rather than strictly following the electronic procedures of corpus linguistics, a mainly qualitative approach is adopted for what is more correctly described as 'corpus work'. The ads were selected in terms of their forming what could be considered a representative cross-section of those present in the issue. Thus the examples provide the reader with a snapshot, albeit a subjective one, of the overall picture. This qualitative approach does however allow some of the original flavour of the selected ads to be conserved and conveyed in the excerpts. The texts were first examined in terms of what seemed significant at an intuitive level and then searched more carefully from a lexico-grammatical viewpoint in order to reveal discursive features and generic patterns. The ads chosen are listed and coded alphabetically in Appendix A. Figure 1 (below) shows one of the ads.

[4] There is also considerable variation in the dimension and layout of the ads.



WANTED

TWO MORE CLASS TEACHERS FOR
HILL MEAD PRIMARY SCHOOL

Foundation Stage/Key Stage One and Lower Key Stage Two

NOR: 474 pupils (Two form entry)

NQTs or RQTs

Are you a Hill Mead teacher?

- You know that Brixton kids can equal the achievement of children anywhere
- You have the ability to make this happen
- You plan creatively and deliver motivating lessons
- You are able to manage behaviour effectively
- You enjoy working with others and you bring out the best in people

REWARD

- A full-time additional, experienced teacher in every year group
- Half a day a week non-contact time guaranteed (a day for NQTs)
- A whiteboard in every classroom and your own laptop
- Good provision for continuing professional development – esp. for NQTs
- A dedicated, hard working team and a good Ofsted
- Tube/bus/trains 5 mins and car parking (outside congestion zone)
- Great community feel
- Cool kids

Come and see us for yourself!

To arrange a visit or to receive an application pack (which can be emailed), call or email:

Hill Mead Primary School, Moorland Rd, Brixton,
London SW9 8UE Tel: (020) 7274 9304
adminofficer@hillmead.lambeth.sch.uk

Closing date 9th May (5:30pm)

Lambeth

Figure 1. Times Educational Supplement job ad (29 April 2007, HMP).

3. Primary school job ads: main components and discursive features

Ads are communicative and interactive. Their rhetorical purpose is to attract and entice. To use the (fishing?) metaphor favoured by the American advertising community, the aim of ads is to ‘hook’ potential customers/clients/applicants. In the case of job ads the purpose is to set in motion a dialogic process between two parties:

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

the employer (in this case the school or LA) and the applicant (here the experienced or newly / recently qualified teacher – NQT/RQT respectively). Indeed the successful completion of the recruitment process will lead to the signing of a binding employment agreement between the two parties.

For advertisers it is obviously important to attract suitable candidates, but as Wignall (2007) points out: “no company wants a flood of applications. It takes time and energy to deal with them”. She then cites an expert in the recruitment advertising sector as observing: “The true skill behind good recruitment is to captivate the right people with the right message” (Wignall 2007). Thus the job ad is an integral part of the selection process and the skill with which an ad is designed and crafted can have significant repercussions on recruitment.

The design and language of ads can be highly fanciful and inventive – they are often extremely successful at catching our attention and tickling our fancy (Cook 1992; Lombardo 2001). This used not to be the case of school ads in England. Twenty-five years ago there was little variety in the ad typology. They tended to be economical in both language and size, only naming the school and the essential details of the post being advertised. These days, however, many schools have decided that they require teachers with a certain kind of profile and attitude and the corpus has quite a number of examples of carefully targeted job ads aimed at singling out suitable applicants, as we will see below.

The ads in the corpus broadly follow a more or less similar generic pattern (Bhatia 2005). Here, I have identified nine basic components, not all of which are always present in the ads, some of which are sometimes grouped together or overlap, and which might be present in a different order.

1. Identifying employer
2. Specifying post(s)
3. Specifying skills and qualities required of applicant
4. Listing and description of school’s qualities
5. Salary and conditions
6. Endorsement(s)

7. Discrimination and security
8. Invitation to visit school
9. Instructions on how to apply

In the subsections below we look at the various components and discursive features of the ads more closely, referring to the schools advertising in the corpus.

Identifying employer

As we would expect the name of the school appears in all the sample ads, sometimes more than once. However, almost all the ads also give both website and email addresses featuring the name of the school – a comparatively recent innovation in advertising, certainly linked to the current omnipresence of the Internet in the social context (Boardman 2005). Thus, for example, the name Hill Mead (HMP) appears four times in the ad in Figure 1.

Many ads also feature logos and advertising slogans. The logo is usually that of the school or that of the LA, like Lambeth in Figure 1. Some schools opt for the more traditional emblems, such as the school coat of arms and motto (perhaps catering to a revival of interest inspired by the success of the Harry Potter books and the highly traditional style of *Hogwart's School of Witchcraft and Wizardry*). The name or the initials might also be incorporated in the logo, or a graphic image associated with the name. Thus Squirrel Hayes First School has a squirrel in its logo (SHF) and St Bernadette's Catholic Primary School (SBC) has its logo in the form of a coat of arms with a cross set between the letters S and B. Christ Church C of E Aided Primary School and Nursery, St Leonard's On Sea (CC) has a large black cross set in a circle formed of the school's motto:

(1) Learning to Live Together/Together Living to Learn (CC)

The names, logos, coats of arms and mottos are unlikely to change unless a school changes radically (for example, as a result of national legislation or local policy), and are thus to be considered as intrinsic parts of a school's identity (Solly 2002). The advertising slogans are of course more specifically linked to a particular ad and thus prone to change; nevertheless they too define and present the school's identity. Those in our sample include the striking slogan used in the Starks Field Primary School ad (SF):

(2) Dream Believe Dare (SF)

The three directive imperatives are then taken up in the main body of the ad: "This is an opportunity to be a part of the DREAM along the way towards excellence", with *DREAM* used as a noun; "You must BELIEVE in our learning values", 'Do you DARE think outside the box and take up this exciting challenge?' (SF, original emphasis). Directives, for Hyland "perhaps the most overt rhetorical strategy of engagement", instruct the reader "to perform an action or to use things in a way determined by the writer" (2004: 22). Here they engage potential applicants, at the same time sending out a clear message as regards the school's identity.

Another strategy used in some of the slogans is the idea of change, often associated with an interrogative, to attract teachers who might perhaps be bored with their current situation. By using questions, the clearest strategy in dialogic involvement, the ads invite their interlocutors to collaborate in reaching an answer (Hyland 2004: 25). Both the following employ catchy questions:

(3) Time for a change? Then come and teach at […] (HGP)

(4) DOES THIS FLOAT YOUR BOAT? (TPS, original emphasis)

The second slogan is only understandable given the fact that the school's name is Tidemill Primary School. Indeed the whole ad plays creatively on the school's riverside name and location on the tidal part of the River Thames in London: the logos of both the school and the London LA of Lewisham are set in stylised black waves at the bottom of the ad, while the top has the eye-catching question in large white capitals on black, which forms a box as it links down the sides to the wavy bottom. The verbal impact of the interrogative is increased by its use of monosyllables and also by its play on the hip '(don't) rock my boat' expression, much in vogue in the world of music.

The ad in Figure 1 also engages would-be applicants dialogically, with its direct question: “Are you a Hill Mead teacher?” (HMP). Its main attention grabbing device is however the classic police ‘wanted’/‘reward’ poster layout. The ‘wanted’ part sets out five essential criteria required of applicants; the ‘reward’ part lists eight potential attractions/rewards. Both parts are highly specific in their references to the profile required of the would-be applicant and to the school, at the same time painting a picture of the school’s identity and setting it clearly within the local context and framework (Canagarajah 2005).

Specifying post(s)

This tends to be a fairly straightforward part of the ads, as we see in the examples below:

(5) We have a number of primary class teacher posts available from September 2007 (HGP)

(6) KS2 Class Teacher […] Required for September 2007 (MCP)

Specifying skills and qualities required of applicant

The ads usually provide a person specification, outlining the qualifications, experience, knowledge and skills required. From a linguistic point of view it is interesting to note the hyperbole of the adjectives chosen to describe the qualities applicants should have. They tend to be evaluative adjectives denoting the highest level of achievement. Thus teachers are required to be “excellent” and “outstanding”, but also “innovative” and “experienced”, as we see below (my emphasis):

(7) **Outstanding** teacher required September 2007 (TPS)

(8) We require an **excellent** Key Stage 1 teacher (HS)

(9) We are looking for an **innovative** teacher (CPS)

(10) We are looking for an **experienced** KS2 teacher (MVC)

(11) We need 3 teachers to join an **outstanding** teaching staff in September (HGP)

Excellent and outstanding are evaluative adjectives typically used by teachers when praising their pupils and thus have a particular resonance in the world of education.

The second person pronoun *YOU* is used a lot in the ads, directly engaging the reader/prospective teacher in the discourse (Hyland 2004: 20), as in Figure 1 (HMP). The first person plural *WE* is also common, again engaging the reader, but also serving to represent the schools not as cold impersonal entities, but rather as bodies made up of children, teachers, the local context. Indeed in the last example above it is the teaching staff which is "outstanding". This shared idea of a community which the successful applicant might wish to join is usually positioned in the description of the school's qualities, which is looked at in the next section.

Listing and description of school's qualities

(12) Tidesmill Primary School is a highly successful, innovative and happy school with a London and national reputation. (TPS)

(13) Churchwood is a school where children discover the champion within themselves. (CPS)

(14) We are a thriving, harmonious inner-city primary school, serving a school rich in social & cultural, religious and ethnic diversity. (MCP)

Again we notice the use of hyperbole to describe the excellence of the schools advertising the posts, as well as the specific reference to the school's diversity in the last of the three examples.

One of the chief ways used by the schools to state their excellence, establishing their credentials (Bhatia 1993: 49-50), is to draw attention to the endorsement of outside bodies. The main one of these is the Office for Standards in Education, Children's Services and Skills (Ofsted), and many of the ads refer specifically (and proudly) to Ofsted in their ads, sometimes directly citing their Ofsted reports, and again we note the use of adjectives – "outstanding" is even used twice in the first example below – as well as the inclusion of the date of the reports: only recent dates are included.

(16) “This outstanding school provides an outstanding education for all its pupils.” Ofsted Jan 2007 (BPS)

(17) Five areas in our recent Ofsted were graded as ‘outstanding’ (LS)

(18) ‘Starks Field is the place to be’ ‘…capacity for further development is outstandingly good.’ (Ofsted 2007) (SF)

(19) “Heseltine Primary School is a good school in which pupils achieve and mature into responsible young people” Ofsted (January 2007) (HP)

Some schools set out their identity through the words of well-known figures in the world of education:

(20)“Imagine that you could become a better teacher just by virtue of being on the staff of a particular school – just from that one fact alone.” Tim Brighouse (GP)

(21) “Too often we give our children answers to remember rather than problems to solve.” Roger Lewin (BHP)

The concepts underlying the two quotations are very different. By quoting Brighouse, one of Britain’s most respected educationalists, the first ad is appealing to the prospective applicant, directly through his imperative instruction “imagine you [...]”, to the concept of the practitioner growing, becoming a better teacher through the quality of the school and staff s/he will join. The opportunities for continuous professional development (CPD) are emphasised in many of the ads, including Figure 1. Lewin, a former editor of the *New Scientist*, on the other hand, places his emphasis on the importance of teaching/learning technique (in this case problem-solving), at the same time sharing the concept with his readers (and thus here with the potential candidates too). Both quotations will certainly have been carefully chosen by the schools to attract suitable applicants to their schools. However, they also reveal much about the schools’ identities, the ethos and values they are promoting.

Some schools display their success by stressing the high numbers of children seeking to enrol, as in the following example.

(22) We are two oversubscribed three-form entry schools. (HS)

In Figure 1 the acronym "NOR" (number on role), refers to the number of children attending the school; "two-form entry" to the number of classes starting out at the school. Some of the language can therefore be considered somewhat specialised discourse, which might be relatively difficult for outsiders to penetrate. This is even more so when the ads refer to salaries, as we will see in the next section.

Salary and conditions

Salaries are set out in language that is likely to be clear to the members of the English educational discourse community, but which could well be difficult for those outside the community to decode, as can be seen in this example.

(23) 2 class teachers (1 MPS [Outer London] and 1 TLR2 £3,941) (SF)

To understand this condensed formulaic discourse (Gotti 2003: 27), they would need to know that teacher salaries in England are based on Main Pay Scale (MPS), which rises incrementally from £19,641 - £28,707 (September 2007 data); that there are special enhanced pay scales for teachers working in or near to London; that experienced classroom teachers undertaking additional responsibility can be given Teaching and Learning Responsibility payments (TLR), which are also scaled.

As well as pay, many schools try to attract teachers by making the working conditions sound attractive, offering incentives to persuade (Bhatia 1993: 52). This is particularly evident in Figure 1, where under the major heading "reward", the school (HMP) emphasises its provision of in-service/(CPD) support for teachers, especially NQTs who are also guaranteed one day a week non-contact time (thus in school but without pupils). The school also offers potential applicants their own laptop as well as an interactive whiteboard in every classroom, and highlights the excellent transport and parking facilities, outside London's expensive congestion zone.

Endorsement(s)

As we have seen, many ads feature the logo of the schools and/or LA. A number of ads also feature logos which reveal information about the school's membership of cultural, sports and other associations and their participation in national or regional schemes in spheres such as ecology and disability. The representation of these logos can be seen both as an implicit statement of identity and also as mutual endorsement. Thus "Green Meadow Primary School, Birmingham" (GM), as well as its own logo centred at the top of the ad and that of Birmingham City Council at the bottom right, has eight others showing its links with bodies such as Sport England and Healthy Schools. Other logos featured in the sample include:

(24) Investor in People (BS / HT / HTPS / GP / GM / MVC / MCP)

(25) Arts Council England (HT)

(26) Eco Schools Award 2006 (BS)

(27) Positive about Disabled People (GM / GP / MVC / MCP)

(28) Dyslexia Friendly (GM / SHF)

Some schools make clear their religious affiliation. For example, the ad for a post at St Bernadette's Catholic Primary School (SBC) is endorsed at the top with the logo of the Roman Catholic Diocese of Westminster and at the bottom by that of the London Borough of Hillingdon. Christ Church C of E Aided Primary School and Nursery, St Leonard's On Sea (CC) has a large cross in its logo and includes the Church of England Diocese of Chichester among its endorsements.

Discrimination and security

It is clear from the ads that anti-discrimination is an important feature in the identity of English schools, reflecting the social and political changes of the last few decades. Indeed, a number of the ads include specific statements of their commitment to anti-discrimination policies. This is very much in step with UK government policy: the current objectives of the Training and Development Agency for Schools (TDA) specifically include the recruiting of more men, people from ethnic minorities and people with disabilities into teaching. A school's commitment to equality is usually expressed in endorsement logos (as we have seen above) or expressly stated in words:

(29) East Sussex County Council is committed to equality of opportunity. (CPS)

(30) We positively welcome applications from all sections of the community. (CPS)

These affirmations could perhaps be seen as mere icing on the cake, reflecting a harmonious and non-discriminatory social fabric. However, the very fact that the schools/LAs include them in the ads probably suggests the contrary. Indeed, inner-city primary schools in multiethnic areas of London such as Lambeth have long been in the front line of the campaign against racial discrimination. Nevertheless the ad in Figure 1 only refers to equality obliquely:

(31) You know that Brixton kids can equal the achievement of kids anywhere (HMP, Figure 1)

The ad expects the reader to know that "Brixton kids" are multiethnic and, by addressing him/her directly, to share the egalitarian concept underpinning the statement. In the second part of the ad the "kids" are referred to again as "cool" and the multiethnicity through the comment that the school has "a great community feel". The implication underscored here, and in many of the ads, is that only those applicants sharing the values and aspirations of the school need apply.

Moreover, a gender issue certainly exists in the English primary school teaching community. According to *Prospects*, the official UK government careers website (available at: www.prospects.ac.uk) a very high proportion of primary school teachers are women. Although as *Prospects* also points out with a somewhat ambiguous comment: "more women now hold senior posts": maybe meaning 'more women than men', but perhaps meaning 'more women now than in the past'. In any case, the ads are carefully worded to avoid any possible charge of gender discrimination.

As regards security and child welfare, recent UK government legislation specifically requires all teachers to have a Criminal Records Bureau (CRB) check and this requirement is often incorporated into the ads, as in these examples, which like many of those in the sample, specifically ask for it to be at the highest level, ‘enhanced’:

(32) The School is committed to safeguarding and promoting the welfare of children and young people and expects all staff to share this commitment. An enhanced CRB check is required for all successful applicants. (GM)

(33) The successful applicant will be required to undertake an enhanced CRB check. (BS)

This CRB requirement is not always present however; for example it is absent in Figure 1 – being a legal requirement, it will be an implicit part of the employment contract. Nevertheless its presence in many of the ads highlights a security issue (and a police/judicial involvement) in English primary schools that was certainly less evident just a few years ago.

Invitation to visit school

The ads use a wide range of different expressions to couch the same basic invitation to visit the school, as we can see below in a list of those used in the sample of just twenty-two ads. This somewhat surprising variety – there is only one instance of exact repetition – could be due to chance or perhaps to the creative linguistic ability of the ad drafters or the wish to differentiate each ad from the others.[5]

(34) Visits welcome (LS)

(35) Visits to the school are encouraged (THPS)

(36) Visits to school are welcome (MVC)

(37) Visits to the school are welcome (BPS)

[5] It needs to be pointed out that the ads were not selected for their different language use on this point, which was only noticed when compiling the list.

(38) Visits to the school are welcomed (CPS)

(39) Visits to the school are warmly welcomed (HT)

(40) Visits to the school are warmly encouraged (CC / GP)

(41) Visits to the school to meet children and staff are very warmly encouraged (SBC)

(42) Visits prior to application are welcome (AJS)

(43) The head teacher actively encourages visits to the school (WPS)

(44) Visits are most welcome and recommended (MCP)

(45) Visits to the schools are warmly welcomed and encouraged – come and see for yourself! (HS)

(46) Interested candidates are encouraged to visit the school (BS)

(47) All interested candidates are positively encouraged to come and visit the school to meet the children and staff and to chat informally […] (SHF)

(48) Visits are by appointment on 3 or 4 May. Please ring if you would like to visit (HGP)

(49) Application packs and appointments to visit can be obtained by calling the school […] (GM)

(50) Anxious to set sail? Then you should call to arrange an informal discussion and request further information (TPS)

(51) If you believe you can add further capacity to our school please do contact the main office on […] to arrange an informal visit with the headteacher / request an application pack (HP)

(52) Please come and look around our school and meet our wonderful children and dedicated staff (BHP)

(53) Please come and visit us for an informal visit – we love showing off! (SF)

(54) Come and see for yourself! (HMP)

Some schools like to restrict the visits to specific dates or by appointment (HGP). Others seem to be always ready for visits, even informal ones (HP/SF). Indeed, Starks Field brazenly admits to "love showing off!" (SF). Hill Mead (HMP) warmly invites those interested to "Come and see for yourself!". Both the last two examples emphasise the warmth and sincerity of their invitation by finishing with exclamation marks. Tidesmill creatively continues its nautical metaphor, with another catchy question: "Anxious to set sail?" (TPS).

Instructions on how to apply

Most schools and LAs these days offer to send applicants an application pack containing: an application form; a job description; a person specification; and information about the school/the school prospectus. This is made clear in the ads, as in Figure 1, where the underlined instruction to "arrange a visit or to receive an application pack (which can be emailed), call or email", then gives the school's street address, telephone number and email address. However, the repetition of "email" suggests that this is the preferred mode and also perhaps the school's intention to appear up-to-date.

4. A word of caution

Print ads are of course a rather special kind of discourse and one where the language presents/is expected to present a partial view in its selling/promotion of a given product/service (Solly 2007). These job ads are no exception and thus the descriptive hyperbole sometimes needs to be taken with a pinch of salt. A humorous take on primary school job ads is given in the 2006 book *How not to teach: Diary of an Urban Primary Teacher*, penned by a Merseyside teacher using the pseudonym Mr Read. He provides an amusing English-English translation of the job description and the reality, including the following (Mr Read 2006):

(55) Idyllic rural location = Miles from anywhere

(56) Dynamic, innovative and creative teacher wanted = The head's a slave driver

(57) NQTs welcome = We've got no money

(58) Improving school = Failed Ofsted

(59) Expanding school = Huge classes

(60) […] would be an advantage = absolutely essential

(61) Challenging inner-city school = Fort Apache – the Bronx

There are some examples of similar language in the sample, for instance as regards behaviour. Green Meadow Primary School is proud of its "responsive and well-behaved children" (GM), but the job seeker would be wise to wonder whether they are always so. Medlock Valley Community School, on the other hand, has "lively and enthusiastic children". They might indeed prove difficult to manage, which is surely why the school is looking for an "an experienced KS2 teacher" who "has high expectations of achievement and behaviour" (MVC). Hill Mead, on the other hand, clearly points out the requirement as one of the five in the applicant's profile:

(62) You are able to manage behaviour effectively (Figure 1, HMP)

The use of the strong present simple form "are able to" leaves no doubt as to the teacher's need to meet the requirement.

5. Conclusion

In its advice to primary school job hunters, the UK Government's guidelines to 'getting a job' (available at: www.prospects.ac.uk) recommend potential applicants keep hold of job ads because they are "usually a good way of identifying the skills and experiences that are most valued by each school/LA," as they provide "real knowledge of the school's outlook, policies and mission." The brief survey of the discourse of a small corpus of English primary school ads presented here shows that the ads do indeed reveal much about the current needs and concerns of the schools, as well as some of their identity traits and those of the social context in which they are embedded.

References

Bhatia, V.K. (1993) *Analysing Genre: Language Use in Professional Settings*, Longman, London.

Bhatia, V.K. (2004) *Worlds of Written Discourse. A Genre-based View*, Continuum, London.

Bhatia, V.K. (2005) "Generic patterns in promotional discourse", in H. Halmari and T. Virtaanen (eds), *Persuasion across Genres: A Linguistic Approach*, Lawrence Erlbaum, Mahwah, New Jersey, pp. 213-228.

Boardman, M. (2005) *The Language of Websites,* Routledge, London.

Canagarajah, A.S. (ed.) (2005) *Reclaiming the Local in Language Policy and Practice*, Lawrence Erlbaum, Mahwah, New Jersey.

Chouliaraki, L. and N. Fairclough (1999) *Discourse in Late Modernity: Rethinking Critical Discourse Analysis*, Edinburgh University Press, Edinburgh.

Cook, G. (1992) *The Discourse of Advertising*, Routledge, London.

Fairclough, N. (1992) *Discourse and Social Change*, Polity Press, Cambridge.

Fairclough, N., G. Cortese and P. Ardizzone (eds) (2007) *Discourse and Contemporary Social Change*, Peter Lang, Bern.

Gotti, M. (2003) *Specialized Discourse: Linguistic Features and Changing Conventions*, Peter Lang, Bern.

Hyland, K. (2004) "Engagement and disciplinarity: the other side of evaluation", in G. Del Lungo Camiciotti and E. Tognini-Bonelli (eds), *Academic Discourse – New Insights into Evaluation*, Peter Lang, Bern, pp. 13-29.

Lombardo, L. (1992) *Selling it and Telling it. A Functional Approach to the Discourse of Print Ads and TV News*, Istituto di Lingue Moderne, LUISS Guido Carli, Roma.

Mr Read, (2006) *How not to Teach: Diary of an Urban Primary Teacher*, Continuum, London.

Pope, R. (2007) "Difficulties in selling schools", *Times Educational Supplement,* 02.03.2007.

Solly, M. (2002) "'Once a trademark, not always a trademark': using language to avoid legal controversy", in M. Gotti, D. Heller and M. Dossena (eds), *Conflict and Negotiation in Specialized Texts*, Peter Lang, Bern, pp. 211-232.

Solly, M. (2007) "'Don't get caught out': pragmatic and discourse features of informational and promotional texts in international healthcare insurance", *Communication and Medicine* 4 (1), pp. 27-5.

Swales, J. (1990) *Genre Analysis: English in Academic and Research Settings*, Cambridge University Press, Cambridge.

Swales, J. (2004) *Research Genres*, Cambridge University Press, New York.

Wignall, A. (2007) "Hire education", *The Guardian*, 13.08.2007.

Appendix A

AJS – Alexandra Junior School, Hounslow
BHP – Bush Hill Park Primary School, Enfield
BPS – Bonner Primary School, Stainsbury
BS – Brookfields School, Reading
CC – Christ Church C of E Aided Primary School, St Leonard's On Sea
CPS – Churchwood C P School, St Leonard's On Sea
GM – Green Meadow Primary School, Selly Oak
GP – George Palmer Primary School. Reading
HMP – Hill Mead Primary School, Lambeth
HGP – Hither Green Primary School, Lewisham
HP – Heseltine Primary School, Lewisham
HS – Hazelwood Schools, Palmers Green
HT – Hounslow Town Primary School
LS – Linchfield C P School, Lincolnshire
MCP – Miles Coverdale Primary School, Shepherd's Bush
MVC – Medlock Valley Community School, Oldham
SBC – St. Bernadette Catholic Primary School, Hillingdon
SF – Starks Field Primary School, Edmonton
SHF – Squirrel Hayes First School, Stoke on Trent
THPS – The Hawthorns Primary School, Wokingham
TPS – Tidemill Primary School, Lewisham
WPS – Woodside C E Primary School, Atherston

2. Corpora in Lexicology and Lexicography

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 121-135 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Remarks on the frequency and phraseology of *A/AN* in Modern English

Stephen Coffey – University of Pisa

1. Introduction

In this paper I look at certain aspects of the word *A* and its prevocalic counterpart *AN*. I will often refer to these two words together as the single item *A/AN*, since the only significant difference between them appears to be their immediate phonetic environment. Specifically, I wish to address the question of what proportion of corpus tokens of *A/AN* are instances of the indefinite article being used as an independent function word. A few examples of such usage are: "Bedrooms have a telephone", "She was holding a candle and her eyes shone in its light", and "There is surely a strong social argument for regarding …".

One reason why *A/AN* may not be performing its role as an independent function word is that it may be carrying a meaning of its own: this happens when it is being used in more or less the same way as the numeral *ONE*. In the three examples above, by contrast, it is probably mistaken to think of *A/AN* as having any meaning of its own. Rather it has been selected to 'accompany' certain words because a certain type of meaning is being conveyed by the text around it; it is reinforcing meaning. It is interesting to note in this respect that if we delete *A/AN* from a longish stretch of text, its omission rarely detracts from communication. Its absence is certainly noted, but above all from the point of view of rhythm.

A second reason why *A/AN* may not be performing its role as an independent function word is that it may be part of a more or less fixed phrase. Sinclair (1999: 160-1) raises this point while discussing the “common words” of English. The example he gives is the phrase *COME TO A HEAD*, in which the various words are co-selected, and the use of *A* cannot really be considered as a textual use of the indefinite article.

This second question is a much vaster issue than the previous one. Many types of phrase are involved, and it is not always clear whether a given group of words is likely to be stored holistically in the mind or not. (For some recent work on the nature and processing of phrasal units, see Wray 2002 and the various contributions to Schmitt (ed.) 2004.)

Corpus methodology

The corpus used in this study was the British National Corpus, World Edition, 2000, hereafter BNC.[1] All examples of usage, including those already given, have been taken from the BNC. Methodology was centred around the analysis of randomly selected individual corpus examples, rather than the automatic retrieval of statistically significant data. There were two main phases of analysis: firstly, the ‘reading’ of sample concordance lines centred around the key word *A* or *AN*; secondly, further corpus searches relating to some specific items noted during the first phase. The corpus was interrogated using the Sara software (see Aston and Burnard 1998).

The concordance lines retrieved in phase 1 consisted of three separate samples of 500 random tokens of *A*, together with another three 500-token samples taken from the spoken sub-corpus. In addition to these, one 500-token sample of *AN* was studied in order to see whether any significant differences emerged. While examining the concordance lines, I was not looking just for repeated patterns: even 500 must be considered a relatively small sample given that there are well over 2 million tokens of *A* classified as an ‘article’ in the BNC, and over 330,000 tokens of *AN*.

[1] For full details, see the BNC website (http://info.ox.ac.uk/bnc/ at the time of going to press).

2. *A/AN* as a content word

In the sample contexts examined, the only times that *A/AN* could be said to be carrying meaning is in its use as an alternative to the numeral *ONE*. The following is an example:

(1) Married with three daughters and a son, John celebrated his fiftieth birthday in February.

An important feature of this sentence is the fact that *A* is comparable with the nearby cardinal number *THREE* (or rather “a son” is comparable with “three daughters”). In cases like this, the extent to which *A/AN* has the feel of a numeral depends very much on the exact wording. In the following example, *A* seems to me slightly less numerical:

(2) Five Tibetan men and a boy of about ten, dressed in worn jackets and trousers, are seated around a fire…

An interesting numerical phrase is the *A* ____ *OR TWO* frame as in “and could have had a goal or two more”. This is often used with units of measurement (e.g. “I mean I’ve known for a week or two”).

The numerical function of *A/AN* is undoubtedly stronger when it is being used in conjunction with a unit of measurement. Two more examples are:

(3) A lovely eighteenth-century house about a mile away.

(4)…which came out just over a year ago.

A/AN is also an integral part of other numbers, and here, too, its function is best viewed as numerical. Both integers and fractions are involved. An example of the former is:

(5) You have to run the gauntlet of a thousand bristling spines.

As regards fractions, *A/AN* is used in different ways. It may be the first element in a ‘proper fraction’ (“with the wars of this century, it had shrunk by a third”), or the second element in a ‘mixed fraction’

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

(“one and a half teaspoons of paprika”); it may also be a constituent of phrases of the type *A THIRD OF* (“in the study by Meshkinpour, over a third of the patients …”).

The types of usage exemplified above link *A* and *AN* to their past since both words are descendants of the numeral now written in the form *ONE* (see *The Oxford English Dictionary*), following what would seem to be an almost universal tendency (Givón 1981: 35). For some further information regarding the use of numerical *A/AN* and its relationship to *ONE*, see Berry (1993: 18-20).

3. Two shillings a yard

The use of *A/AN* described in this section is a rather anomalous one. It may be exemplified by the following sentences:

(6) You can’t live on 40 pounds a week.

(7) There was cotton gingham at two shillings a yard.

Given the fact that *A/AN* is used here directly before a unit of measurement, and given the nearby (preceding) presence of a number (e.g. “40 pounds”), it would be natural to think in terms of the numerical use of *A/AN*. Paradoxically, however, it is not possible to substitute *A/AN* with *ONE* in this structure. Nor does it make much sense to think of *A/AN* as a number here, since no other number is possible, and numbers are normally the result of paradigmatic choice.

The *raison d’être* for this unusual type of phrase is the fact that *A/AN*, in this usage, was originally a preposition (see OED: a, a^{2} [indefinite article], usage no. 4).

In modern English the usual structure of this frame may be schematized as ‘number + noun group + *A/AN* + unit of measurement’; occasionally a quantifier is found as the first element instead of a number (e.g. “I walk the dog up there several times a week”), and the words *ONCE* and *TWICE* may take the place of the first two slots together (e.g. “letters were delivered twice a week only”). As can be seen, the frame is a fairly open one, lexically speaking, with *A/AN* being the only fixed component.

4. *A/AN* in phrases which modify or complete a noun or noun group

A number of phrases containing *A/AN* serve to modify a noun in some way. Just a few of these follow the noun: examples are *OF A SORT* ("there are, however, clues of a sort") and *AS A WHOLE* ("and of the economy as a whole"). The majority precede the noun: I now illustrate and discuss the more frequent or more interesting cases.

A first set of pre-modifying phrases are the quantifiers *A FEW*, *A LITTLE* ("just a little milk in my tea"), *A GREAT MANY*, and *A GOOD MANY*. To these we may add the syntactically unusual *MANY A* ("we have had many a grave difference on questions of policy").

Next there are phrases which are realizations of the *A/AN* _____ *OF* frame, (or to be more accurate, the *A/AN* _____ *OF* _____ frame). Quite a number of these are quantifiers, though there are other meanings involved as well. Examples of phrases found in the corpus samples are: "a bit of", "a little bit of", "a great deal of", "a lot of", "a number of", "a quantity of", "a couple of", "a handful of", "a drop of" (e.g. port), "a load of" (e.g. rubbish), "a total of", "a matter of" (e.g. judgement) and "a kind of".

In the case of *A/AN* _____ *OF*, it is not always clear whether or not we should be talking in terms of a fixed phrase, and this is perhaps destined to remain the case until we have a better knowledge of how language components are stored in the mind. An example of a phrase which leaves me undecided is *A MIXTURE OF* in the frame '*A MIXTURE OF* + plural noun' (e.g. "precipitated by a mixture of factors").[2] For more on lexico-grammatical frames involving high-frequency function words, including *A*_____ *OF* and *AN* _____ *OF*, see Renouf and Sinclair (1991).

The remaining phrasal types I discuss in this section all have one thing in common: they have *A/AN* as their final component. An example we have already seen is the phrase *MANY A*, as in "many a grave difference". There are a number of other phrases of this type, and in all cases I prefer to view the phrase as a paradigmatic

[2] This frame is different from, and much less common than, the frame '*A MIXTURE OF* noun (+ noun) *AND* noun' (e.g. "a mixture of English and Continental"; "a mixture of anger, fear and pleasure").

alternative to *A/AN*, rather than as the indefinite article preceded by a 'pre-determiner' of some sort. Using the example just given, the phrase is therefore to be seen as "*MANY-A* + *GRAVE* + *DIFFERENCE*", rather than "*MANY* + *A* + *GRAVE* + *DIFFERENCE*". This puts *MANY A* on a par with, for example, *A FEW*.

A first group of phrases of this type belong to the frame *A/AN* _____ *OF A/AN*_____ . An example is "It's a hell of a coincidence". Here, there are two instances of *A*, though in this particular phrase the first *A* is sometimes replaced by *ONE* ("a vivacious personality and one hell of a figure") and occasionally omitted ("must be hell of a shock being told"). Note too the informal spelling *A HELLUVA* ("you gave me a helluva fright"), which reinforces the notion that the second *A* belongs to the pre-nominal phrase. Some other phrases which fit into this frame are: *A HECK OF A/AN*, *A DEVIL OF A/AN*, *A DICKENS OF A/AN*, *A BUGGER OF A/AN*, *A SOD OF A/AN*, *A CRACKER OF A/AN*, and *A BIT OF A/AN* (as in, "I just felt it was going to be a bit of a funny day"). Note, too, the fixed phrase *A WHALE OF A TIME*.

Another two pre-modifying phrases I would place in this category are *SOMETHING OF A/AN* ("he had caused something of a furore") and *SOMEWHAT OF A/AN* ("it too became somewhat of a curiosity"). Then there are the frames *QUITE A/AN* _____ ("that was quite a good game actually" and *RATHER A/AN* _____ ("it is rather a complex and grey area"). In the case of *QUITE*, an additional argument in favour of the *QUITE A/AN* _____ frame (as opposed to *QUITE* + *A/AN* + noun) is the fact that a mere 8 phrases account for over 48% of all corpus tokens (*QUITE A LOT*, *QUITE A FEW*, *QUITE A BIT*, *QUITE A WHILE*, *QUITE A NUMBER (OF)*, *QUITE A LONG TIME*, *QUITE A LONG WAY* and *QUITE A SHOCK*). It seems reasonable to suppose that when very frequent phrases are stored in the mind as units (as seems probable with at least some of the phrases listed above), then this fact may help the frame to be stored as well, ready to be completed by other, less frequent words.

Two other words customarily listed alongside *QUITE* and *RATHER* as 'pre-determiners' are *WHAT* and *SUCH*. Again, I would prefer to view these as contributing to the frames *WHAT A/AN* _____ and *SUCH A/AN* _____ . Some examples of usage are:

(8) What a performance!

(9) What a change in the boy, sir!

(10) It's such a pretty name.

(11) It's such a pointless, sad way to die.

The latter two examples are with *SUCH A/AN* being used as an intensifier. We could also extend this phraseological view of *SUCH A/AN* to its anaphoric use (e.g. "many people found such a prospect disagreeable", "such a thing is very common").

A final frame (or set of frames) I will mention may be schematized as '*HOW* / *SO* / *TOO* + adj + *A/AN* + noun', where the key words *HOW*, *SO* and *TOO* introduce a syntactically unusual expression. Some contextualized examples are:

(12) Rather complex for so short a work.

(13) I was amazed at how big a sound was coming back.

(14) We're not having too early a lunch, are we?"

Some specific adjectives and nouns tend to be used together in this frame. For example, of 49 tokens of *SO SHORT A*, 31 are completed with the word *TIME*, and of 41 relevant tokens of *SO* ___ *A TIME*, 31 are completed with *SHORT*.

5. *A/AN* in other expressions

Section 4 was about pre-nominal phrases. I turn now to other lexico-phraseological environments in which *A/AN* is often found. The grouping of phrases below is totally pragmatic: the first two sections bring together items which have syntactic similarities, the third is a semantic grouping, and the fourth contains examples not included in any other categories described.

Verbal expressions

A first set of verbal phrases consists of conventionalized expressions which involve some sort of underlying metaphor or simile. Examples are "to put a damper on", "to give someone a free hand", "to have come a long way", "to take someone for a ride", "to turn a blind eye" and "not to give a hoot".

A second set are verbal expressions of a less idiomatic nature, often incorporating very common verbs. Some contextualized examples are:

(15) I'll go and have a bath then.

(16) Not that I make a fuss about it.

(17) He can have a chat wi' you next time he comes.

(18) I will give her a ring in a few minutes.

(19) You stand by the gun and give me a shout in two hours' time.

Some further, non-contextualized examples are *TO COME AS A SURPRISE*, *TO HAVE A VESTED INTEREST IN*, *TO GO FOR A WALK* and *TO TAKE A FRESH LOOK AT*.

Sometimes it may be less obvious that we are dealing with a verbal construction because the verb is a link verb, and is therefore less salient. The following is an example involving the word *DODDLE*:

(20) I was heard to remark that my job would be a doddle.

Corpus data suggests that *DODDLE* is almost always preceded by *A*, and that *A DODDLE* is almost always preceded by a form of the verb *TO BE*. It would be reasonable, therefore, to talk in terms of the expression *TO BE A DODDLE*.

At this point it is worth mentioning the general phenomenon whereby some English nouns, in some of their uses, are typically preceded by *A/AN*. Various semantic groupings of such nouns (including *DODDLE*) are to be found in Francis *et al.* (1998: 37-41). Where the noun in question is typically used with a certain verb (or in some other typical structure), then it may be useful to consider it as a fixed phrase.

Functional phrases beginning with a preposition

Another commonly found category consists of a disparate set of phrases beginning with a preposition, and sometimes including slots to be filled, optionally or obligatorily. There are a number of meanings and functions involved, and I will limit myself to listing a variety of the phrases found: “with a view to”, “for a reason”, “as a result”, “as a result of”, “to a [remarkable] degree”, “to a [certain] extent”, “on a [grand] scale”, “up to a point”, “from a [financial] perspective”, “[more than one thing] at a time”, “in a row” (= consecutively), “as a basis for”, “on a [regular] basis”, “for a change”, “for a start”, “in a way”, “in such a way”, “in a sense”, “as a matter of fact”, “at a guess”, “at a glance”, “at an [alarming] rate”.

A/AN *in expressions of time*

The next group of phrases is that of TIME expressions. Some examples from the corpus samples are: “for a long time”, “for a bit”, “for a moment”, “a moment later”, “a split second later”, “a short time ago”, “a moment or two ago”, “wait a minute”, “hold on a minute”, “in a second”, “in a minute” and “in a while”. Phrases such as these may be grouped together in various ways according to their component parts, thus forming a number of variable frames, for example *IN A* ____, *A* ____ *LATER*, *FOR A* ____.

A medley of items

I complete this overview of phrasal uses of *A/AN* by bringing together other types of phrase found in the corpus samples and not yet accounted for. There are figurative adverbial phrases (“like a shot”, “like a brick”, “in a flash”), clause- and sentence-level items (“if that isn’t a rude question”, “there’s a surprise”), one well-known song title (“it’s been a hard day’s night”), and one structurally independent noun phrase (“half a crown”). There are also many phrases containing very common words mentioned already (*BIT*, *LITTLE*, *LOT*, etc) but which are employed in a different way to those mentioned in section 4: there are phrases which modify adjectives or adverbs (“they tend to become a bit intolerant of others”), phrases which modify verbs (“we’ll work a little bit on the maths”, “we deal with the press a lot”) and phrases which are pronominal in nature (“Jimmy Connors inherited a lot to begin with”).

6. Incomplete *A/AN* …

Although it is customary to think of *A/AN* as being used in conjunction with a following noun, this is not always the case. In spoken language we often interrupt what we are beginning to say and reformulate our discourse ("is it a that like moves around on er wheels?"), or else we break off to listen to what someone else is saying. Also, *A* and *AN* are sometimes repeated while the speaker hesitates ("I was going to bring a, a poster for you and I've forgot it"), and it could be argued that in such cases only the final *A/AN* is an actual realization of the indefinite article.

With regard to repetition in particular, there are over 2000 instances of "a a" in the BNC, mostly in the transcribed spoken texts. Where these occur at the beginning of a noun phrase, both *A'*s are usually tagged as 'articles'. There are also over 1300 "a, a" (that is, with an intervening comma) in the spoken corpus transcriptions, again typically both identified as articles. There can also be much longer strings (e.g. "the view we've taken is that it involves a, a, a, a, a great deal of administration"). And there are also switches from *A* to *AN* (e.g. "and that's a, an incredible tale"), as well as repeated *AN*s ("we had what I would term an an outside trade"). Some "an an"s lead on to the word *AND*, and in this case they tend to be classified as conjunctions (or else are 'unclassified').

There are also slightly richer, lexically speaking, repetitions ("erm I think perhaps it was a good move in a, in a, in a way", "And the whole street was revitalized in a in a in a swoop."). For more on repetition of this sort and uncompleted utterances, see Biber *et al.* (1999: 1055-1064), Brazil (1995: 211-212) and Sinclair and Mauranen (2006: 80).

7. *A/AN* and frequency

The question posed at the beginning of this article was: "what proportion of corpus tokens of *A/AN* are instances of the indefinite article being used as an independent function word?". The answer to this question is neither precise nor straightforward. It cannot be precise because automatic analysis could not possibly return valid results at the present time, and analysis of sample data will perforce be approximate in nature.

There cannot be a straightforward answer to the question since there are too many variables. Do we count all the repeated tokens of *A/AN* as instances of the indefinite article, and the various tokens left hanging, when the text changes direction? Do we count the more or less numerical uses, such as “three daughters and a son” and “a mile away”? Should we, as I have suggested, consider *QUITE A* ___, and several other pre-modifying phrases, as holistically stored frames? Is the *A/AN* in “two shillings a yard” to be considered as the indefinite article, and, if so, is the relative frame stored holistically? In the frequency data below, I have tried to obviate problems such as these by breaking the data down, as will be seen in Table 1.

Frequency data

In this section I present frequency data regarding the word *A* tagged as an ‘article’ in the BNC. I do not at present have comparable data for the word *AN*. The information relates to the three 500-token samples taken from the whole corpus and the three taken from the Spoken corpus (which comprises about a tenth of the whole corpus). I refer to these samples as: BNC-1, BNC-2, BNC-3, Spkn-1, Spkn-2 and Spkn-3.

The data is summarized in Table 1, and is organized around 10 information types:

1. ‘tagging errors’;
2. ‘unclear’: it is unclear to me what is happening in the text;
3. ‘incompletion’: the noun phrase is not completed, or is completed after repetition of an ensuing *A*.

Rows 4 to 6 have been included as separate data since some readers may prefer to consider some or all of the phrases as ‘normal’ instances of the indefinite article. In this case, the figures should be integrated with those in row 10.

4. ‘numerical use with unit of measurement’: *A MILE*, *A MONTH*, etc. It is to be noted that in the samples discussed here, there were no examples of numerical *A* with other nouns (e.g. “three daughters and a son”), and for that reason there is no row in the table devoted to this information;
5. ‘rates’: “two shillings a yard”, etc.;
6. ‘phrases ending a’ (certain of the pre-modifying frames discussed in section 4): [something/somewhat] of a, a [heck,

etc] of a, [quite / rather / such / what] a, [how / so / too] [short] a.

Rows 7 and 8 have been included as separate data since these phrasal types show the biggest differences between the whole corpus and the spoken component.

7. *A BIT*, *A LOT* etc.: I have here grouped together the following very frequent items: *A BIT*, *A FEW*, *A LITTLE*, *A LITTLE BIT* and *A LOT*, taking various syntactic roles, and with or without following *OF*.
8. 'within numbers', e.g. *A THOUSAND*, *ONE AND A HALF*.
9. 'other lexical phrases': all other phraseological items considered to be highly lexicalized.
10. 'more independent function word': uses of *A* as a more or less independent function word. The "more or less" is of significance here since there are different degrees of independence, ranging from the very open ("the water being collected in a cistern beneath") to more familiar sounding phrases ("in a black floppy hat", "make a significant impact on", "according to a recent survey", "walks with a limp").

		BNC-1	BNC-2	BNC-3	Spkn-1	Spkn-2	Spkn-3
1	**Tagging errors**	1	-	-	2	1	2
2	**Unclear**	1	-	-	2	1	1
3	**Incompletion**	1	-	3	30	24	24
4	**Numerical use with unit of measurement**	3	2	4	5	13	14
5	**Rates**	4	1	4	8	4	5
6	**Phrases ending *A***	6	13	4	5	20	8
7	***A BIT*, *A LOT*, etc**	11	21	24	56	65	48
8	**Within numbers**	3	4	-	11	12	15
9	**Other lexical phrases**	59	54	48	66	68	61
10	**More independent function word**	411	405	413	315	292	322
	Total	500	500	500	500	500	500

Table 1. Frequency data regarding *A* in the British National Corpus.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

As is immediately apparent from this data, there is a notable difference between the samples from the whole corpus and those from the spoken corpus as regards the number of tokens functioning as a more or less independent function word. There are three main reasons for this difference: firstly, the quite frequent repeats and incompletions found in the spoken language (row 3); secondly, the much higher instance of phrases such as *A BIT*, *A LITTLE* and *A LOT* in the spoken corpus (row 7); thirdly, the higher presence in the spoken corpus of *A* being used within numbers (row 8).

The third of these differences is partly related to the way in which numbers are transcribed in the spoken corpus. Consider as an example the number which is sometimes indicated in Latin by the letter *C*, and which usually appears in written English as *100*, *A HUNDRED* or *ONE HUNDRED.* Ignoring the latter (not strictly relevant to the present discussion), the relative frequencies of *100* and *A HUNDRED* in the written (WR) and spoken (SP) parts of the corpus are: *100* (WR 6889, SP 19), *A HUNDRED* (WR 2148, SP 2573). That is, the spoken transcriptions much prefer the alphabetical option. The same tendency was found for a few other numbers sampled.

Table 2 shows the proportion of instances of *A* used as an independent function word, that is, corresponding to the figures in row 10 of Table 1. Row 1 shows the proportion in relation to each complete 500-token sample. In row 2, there is a very slight adjustment, since I have excluded *a priori* any cases of incorrect tagging or of contexts which were not clear to me (so, for example, in row 2 the figure for BNC-1 is based on 498 tokens). The third row shows what happens if we choose to further exclude from our calculations all instances of incompletion, or of completion after an ensuing *A* (corresponding to row 3 in Table 1).

Corpus sample	BNC-1	BNC-2	BNC-3	Spkn-1	Spkn-2	Spkn-3
All 500 tokens	82.2%	81%	82.6%	63%	58.4%	64.4%
excluding errors and unclear instances	82.5%	81%	82.6%	63.5%	58.6%	64.8%
excluding errors, unclear instances, and incompletion	82.7%	81%	83.1%	67.6%	61.6%	68.1%

Table 2. Proportion of tokens of *A* used as an independent function word.

The overall figures for *A* as an independent function word in the three ‘BNC’ samples are quite similar, and this suggests that 500 is probably a big enough figure to give a general idea of what is going on in the corpus as a whole, this despite the fact that each sample corresponds to only about 0.025% of the whole. In the spoken samples, there is an evident difference (though by no means enormous) between the second sample and the other two; this suggests that bigger sampling may result in more reliable data.

Conclusion

As part of language description, all words need to be described as fully as possible. In the case of the very high-frequency words of English, it is important that the description of the individual items be as complete as possible, since they do not necessarily lend themselves to a great deal of shared description with other items of the lexicon (see Sinclair 1999).

In this article I have attempted to come to grips with one aspect of the use of the indefinite article: the frequency with which it occurs as an independent function word in text. The question is quite a complex one, for the various reasons mentioned in the course of the article. It is also a ‘technically’ difficult one, especially when one’s starting point is a large corpus. I hope to have given answers which are at least in the correct order of magnitude.

References

Aston, G. and L. Burnard (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh.

Berry, R. (1993) *Collins Cobuild English Guides: 3, Articles*, Harper Collins Publishers, London.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *The Longman Grammar of Spoken and Written English*, Longman, Harlow.

Brazil, D. (1995) *A Grammar of Speech*, Oxford University Press, Oxford.

Francis, G., S. Hunston and E. Manning (1998) *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*, Harper Collins Publishers, London.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

Givón, T. (1981) "On the development of the numeral 'one' as an indefinite marker", *Folia Linguistica Historica* II (1), pp. 35-53.

The Oxford English Dictionary Online, Oxford University Press, Oxford. http://dictionary.oed.com/

Renouf, A. and J. Sinclair (1991) "Collocational frameworks in English", in K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics: Studies in Honour of Jan Svarvik*, Longman, London, pp. 128-143.

Schmitt, N. (ed.) (2004) *Formulaic Sequences: Acquisition, Processing, Use*, John Benjamins, Amsterdam.

Sinclair, J. (1999) "A way with common words", in H. Hasselgård and S. Oksefjell (eds), *Out of Corpora: Studies in Honour of Stig Johansson*, Rodopi, Amsterdam, pp. 157-179.

Sinclair, J. and A. Mauranen (2006) *Linear Unit Grammar*, John Benjamins, Amsterdam.

Wray, A. (2002) *Formulaic Language and the Lexicon*, Cambridge University Press, Cambridge.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 137-152 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Disambiguation of English pre- and postmodified noun phrases

Valerio Fissore and Ruth Anne Henderson
University of Turin

1. Introduction

This essay is the fruit of the very different, yet complementary, interests of the two signatories. The first part presents preliminary considerations regarding the use of noun phrase premodification in English and certain difficulties of gifted non-native speakers in its application. The second part looks at both pre- and postmodification of the noun phrase, with special reference to scientific and medical texts and to the problem of disambiguation, bearing in mind, even though not discussing, the increasing use in scientific fields of automatic translation devices.

The marked tendency in English to premodify a noun phrase by means of another noun phrase is one which can perhaps be detected as early as the Old English period, in which the kenning, a metaphorical compound noun composed, in fact, of noun + noun, was a widely used poetic device: thus *hwælrad* (whale-road, sea), *merehengest* (sea-steed, ship), *sigedryhten* (victory lord, God) and many more. What is noticeable here is the inevitable use, as first element of the compound, of an uninflected singular substantive – exactly as in both modern compounds such as *SHOELACE*, *BIRTHDAY*, *CARPORT*, *MOUSEPAD*, and in two-word fixed expressions like *BATHROOM WINDOW*, *PENCIL SHARPENER*, *PHILOSOPHY COURSE*. The pattern, then, would appear to be native to English, and is, indeed, a feature of the Germanic languages in general.

2. Noun phrase premodification

The corpus on which the following observations are based is of 53 essays by an unusually gifted class of second-year University students,[1] chosen because their linguistic ability means that they rarely made elementary errors; hence, the imperfections found in their work represent difficulties not uncommonly encountered in the use of noun-phrase premodification, even at a high level.

For foreign learners, especially at upper intermediate or advanced level, the apparently ubiquitous application in English of premodification in noun phrases represents something of a mixed blessing. On the one hand, this structural device offers the possibility to economise on sentence length (and Italian students in particular notice and often admire the relative brevity of the English period); on the other hand, it is by no means simple to manipulate. Superficially, for example, there is no significant difference between the postmodifying prepositional phrase in noun phrases such as *THE WINDOW OF THE BATHROOM* and *THE BEGINNING OF THE CHAPTER:* in Chomskyan terms, their surface structure is identical. Yet *THE BATHROOM WINDOW* is a valid alternative to *THE WINDOW OF THE BATHROOM*, whereas **EVERY LINE BEGINNING* is unacceptable. Premodifiers of this second kind are not infrequent in the written work of students whose command of English is generally excellent: thus, in a recent essay, **WOMEN HARASSMENT*. There is no obvious explanation for the exclusion of such phrases; simply to say that they "sound wrong", though true, is unhelpful. The native speaker would appear instinctively to limit this kind of premodification to locations (*THE TOWN CENTRE, THE GARDEN GATE*) and concrete objects (*THE WARDROBE DOOR, A PENCIL SHARPENER*), or to abstract concepts (*A PHILOSOPHY COURSE, A LOVE SONG, A POLICY MEETING*), adopting a different construction for references to human beings (*THE HARASSMENT OF WOMEN → WOMEN'S HARASSMENT; A BOOK FOR CHILDREN → A CHILDREN'S BOOK*).

Here again, however, English lays traps for the unwary. The last example above, *A CHILDREN'S BOOK*, has the singular indefinite article *A* immediately preceding the plural Saxon genitive *CHILDREN'S*, though the article refers to *BOOK*; yet **A SHAKESPEARE'S SONNET* or **THE FORSYTE SAGA SUMMARY* calls for the red pen. In this latter case,

[1] Including a Hungarian, a Turk, an Iranian, two Rumanians and an Albanian.

the proper noun has to be treated as an adjective, hence stripped of genitive inflection, lest *A* be perceived as referring not to *SONNET*, but to *SHAKESPEARE.* In this case, insofar as a rule of thumb can be proposed, it can only be “Never use the definite or indefinite article before a personal proper name unless that name is adjectivised”.

Many of the errors typical of non-native speakers, and not instantly accessible to disambiguation, involve the substitution for the expected prepositional phrase of a premodifying NP: thus **A WORD EXPLANATION* (the explanation of a word), **THE BOAT OARS* (the oars of the boat), **A STASIS FEELING* (a feeling of stasis), **THE MOON ECLIPSE* (the eclipse of the moon), **BLOOD CIRCULATION* (circulation of the blood), **THE SIN TOPIC* (the topic of sin). Note that a) in every one of these cases (from the corpus indicated above), the solution lies in the use of a PP with 'of' as the preposition; b) in every case but the first (*A WORD EXPLANATION*) the use or absence of the article has been based on the Head of the NP, and not on the premodifying noun.

A further subtlety of premodification is the convention whereby, when the head of a noun phrase is preceded by two or more adjectives, they are listed paratactically: *A TALL, HANDSOME, BEARDED MAN*, *THE FIRST FINE CARELESS RAPTURE* (from Browning’s “Home Thoughts from Abroad”), except where there is tension or direct contrast between the adjectives in question: *A SHY BUT COURAGEOUS CHILD*, or where one of the adjectives is derived from a verb participle: *A DOMINATING AND JEALOUS FATHER* (where, in addition to the difference of form, the two premodifiers refer respectively to outward-directed and interior characteristics). Here again, the foreign learner tends to insert *AND* in every case before the last adjective: thus *PERSONAL AND DEEP EMOTIONS*, *A LYRICAL AND PASSIONATE TONE*. Though it would be an exaggeration to define these as errors, they are nevertheless not the form which would occur spontaneously to the native speaker.

Other subtleties may be lost on foreign learners, even at advanced level. One of these is certainly the partial restriction of premodifying adjectives to permanent or at least enduring situations, as indicated by Quirk and Greenbaum (1973: 377): “[…] those adjectives which cannot premodify have a notably temporary reference”. A typical error in this respect is found in a comment on the lines in Browning’s “Meeting at Night”:

And the startled little waves that leap
In fiery ringlets from their sleep,

where one student wrote of “the asleep waves”. However, Quirk and Greenbaum’s first affirmation is not reversible: it is not true to say that “those adjectives which have a notably temporary reference cannot premodify”, as examples such as “the sleeping waves”, “the barking dog”, “the frightened child” make clear. It is therefore not the case, as Quirk and Greenbaum say (1973: 400) that “Premodification confers relative permanence”. (The wording of this second affirmation is odd, seeming as it does to suggest that the idea of permanence resides not in the semantic content of the phrase but merely of the word order: i.e. that the meaning, the definition of the individual words is of less significance than their position in the phrase).

3. Noun phrase complexity in scientific discourse

All text-types and texts may be defined in terms of the complex role they have been designed to perform in human society. This role can be called their sociolinguistic function. The sociolinguistic function of a given text is identified and pursued before the text is produced and used, and should be identified and taken into consideration before any linguistic investigation be done on the text; descriptive-interpretive hypotheses about it may/should be put forward to guide investigation (‘may’: because we already know something of the scope of a communicative text or text-type well before we experience it; 'should': for the researcher not to be overwhelmed – by doing a blind inquiry – with information for which s/he may have no use).

No reference to one ‘corpus’ of texts will be made in this section, but a variety of sources will be used: specifically a selection of contributions to medical journals, *The Lancet*, *The British Medical Journal*, and the *Proceedings of the National Academy of Sciences* (PNAS) of the United States, years 1999-2001; other texts drawn from a miscellaneous bibliography. The ‘corpus’ has been used as a quarry, out of which examples have been fished to prompt hypotheses to put to the test. For a corpus to be a useful quarry for research, one must know what to look for in it; otherwise, one is likely to get lost in a wasteland or a desert. Strategic choices must be made before the corpus is tackled at all.

Let us begin with a few preliminary considerations on how (the English) language is used in the formulation of scientific discourse. The language of science has been selected as specifically appropriate for the discussion intended in this paper, about the syntax and disambiguation of English NPs, because of its very special complexity in this respect, which does not contradict ordinary use, even if it represents an extraordinary application of it.

The language of scientific and technical discourse is characterised by a mathematical affiliation with the content of its topics, and it could not be otherwise: words in scientific discourse tend to be used as, or instead of, mathematical symbols – very much richer than mathematical symbols, in fact, and for this reason sometimes intractable, but the combination of them to make meaning is very much the same. This holds true whatever the scientific topic; not necessarily mathematics or any topic obviously closely related to mathematics, such as physics, or chemistry, or the engineering of bridges. All discourse aspiring to scientific status must conform, or at least must tend to conform to it. The nouns, or nominal groups, are the figures; the verbs are the mathematical operators (or rather the verbs 'tend' to perform in syntax like mathematical operators). This fact is conspicuously apparent in the following examples from the texts of the selected scientific quarry:

(1) These diets increase blood cholesterol, which is related to risk of coronary disease.
"Increase" implies both addition and sum total, cholesterol on cholesterol.
"Is related to risk" implies causality, "leads to", and has therefore the value of =.

(2) 3,4-methylenedyoxymethamphetamine MDMA (ecstasy) is an amphetamine congener that has gained popularity as a recreational drug.

Leaving the moral comment aside, if the acronym MDMA is exploded, the result of the explosion is a mathematical addition: methylenedyoxymethamphemine, the chemical compound that is given that name. "Is" ("*is* an amphetimine congener") that follows also functions as the mathematical sign =.

Even a sentence like the following:

(3) The obstetric nurses change shifts at 7 am, and the nurses have full responsibilities for each delivery.

which is descriptive of hospital procedures, is like arithmetic in that it describes a recommended sequence of actions for safe clinical results. Note the repetition of "the nurses": there are no synonyms, for the words tend to the status of figures in this type of discourse.

The small sample may appear to be too small for permitting hypotheses about how scientific discourse is organised; it is economical, however, and it is undeniable that some traits may be sufficient to impose the fact of their recurrence even only to a cursory reading; the recurrence will alert the researcher towards speeding up closer observation of probably relevant facts; for example, even a naive reader would observe that the verb *BE* is always there both as a copula and as an operator for the passive voice; that *HAVE* is there twice (in argumentation *BE* and *HAVE* are the only inevitable operators; any physical or mental object may be defined in terms of what it is permanently (*BE*) and what is contextually accessory to it (*HAVE*): a tree is a tree, and may be small or big, living or dead, useful or useless i.e., it has size, + life or – life, + leaves or – leaves, usefulness or uselessness to characterise it).

About *INCREASE* in (1), it has already been noted that it implies addition; about *GAIN* in (2), we note that it has distributive value, and *CHANGE* in (3) implies commutation and substitution. Relational verbs delimit the semantic span of the verbs used in ESP: verbs may be assumed to be mainly in the relational role.

It may be further observed:

- that sentence structure does not exceed the two clauses, and is therefore very economical (of course sentences with a larger number of clauses are to be expected, but the reader is contextually already invited to assume that this will not be the rule);
- that there is use of (different degrees of) specific lexis. Things must be called by their names;
- that the constituent noun phrases tend to favour a compound/ complex structure (“blood cholesterol”; “coronary disease”;

"3,4 methylenedyoxymethamphetamine MDMA (ecstasy)"; "amphetamine congener that has gained popularity as a recreational drug"; "recreational drug"; "obstetric nurses");

- that juxtaposition, though here testified only in one of the three examples, suggests itself as likely to be a frequent syntactic feature in a text type aiming at precision combined with economy. Juxtaposition is typically the replica of real life: connections other than spatial or temporal proximity are not visible features, they are established by the human mind; in scientific discourse connections are often left unstated, objects or events are seen physically one beside the other when there is assumed certainty that they are received as taken for granted, a matter of fact.

Now the focus will be placed on one of the linguistic features of scientific texts: the noun phrase. The nominal syntagmatic unit may be expected to carry the weight of scientific observation and discussion, along with terminology, and is very often indistinguishable from it (typically, glossaries of scientific terminology contain a high ever increasing number of multi-word entries): the noun phrase – the forms it may take – must, by necessity, perform an important role in a text-type characterised by a dominant ideational function and a pragmatic communicative goal.
J.C. Sager *et al.* write:

> The most important components of the vast majority of SE [Special English] sentences are conceptual units expressed in nominal groups. They contain the individual items of information which make up the detailed description of a machine or process, the logical exposition of an idea or theory, the reasoned explanation of natural phenomena and the objective evaluation of experimental data. They act as the building blocks from which SE sentences are constructed because they possess certain inherent qualities which enable them to perform the task of communicating information effectively and efficiently, namely:
>
> - they can be placed as subject or object and then emphatically within syntactic constraints at the beginning or end of the sentence according to their relative importance;
>
> - they can be combined by means of structural words such as connective verbs, conjunctions, prepositions and relative pronouns;

> - their information content can be expanded by the insertion of different types of modifiers.
>
> The possibilities of combining, extending and sequencing nominal groups available to the specialist writer are limited only by the ability of the intended recipient easily to comprehend the resultant text (Sager *et al*. 1980: 219)

This description aptly confirms the linguistic quality of the NP to state its identity arithmetically, via the addition of meaningful algorithmic units. A word may be ambiguous, but is made less so by the addition of other words which state features that establish the limits according to which that word is to be understood: *MAGNETIC RESONANCE IMAGING*, in the medical context, describes a procedure whose referent in real life is unambiguous, while each of the constituent words is by itself more difficult to pinpoint. *IMAGING* is a recently coined technical term (but in everyday use of English, or in literature, it could develop/have developed different ambiguous applications) which helps to restrict the implications of its modifiers.

Sentences are made up of combinations of NPs and VPs, and their own several combinations. Following what was said above on the likelihood for the verb to tend to being either *BE* or *HAVE*, we are little surprised at the generally maintained observation that the VP, also in terms of semantic, tense, and aspect quality, shows little variation. In contrast, the noun phrase, as the focus of scientific discourse, is precise but at the same time inevitably shape-shifting, protean. No other functional phrase can be, and is, manipulated to the extent that the NP is. Besides, English language realisations of scientific discourse terminology can rely on the advantages of a greatly varied flexibility.

The English NP is characterised by the possibility of massively and variously modifying its head, more or less indefinitely[2] on the left hand side and on the right hand side.

Left-hand-side modification (pre-modification) is characterised by juxtaposition and sequence. The left- hand-side 'slots' are occupied by deputed classes of words. Leaving aside the slots occupied by predeterminers, determiners, numerals which collocate

[2] 'Indefinitely', of course, is used only for emphasis. Sager *et al.* (1980) say the same more sedately.

in obligatory reciprocal connections, it is important to consider the variety of epithets that can be inserted between these and the head. Epithets also have their obligatory position, which is determined by the nature of the epithet, even though it is not uncommon for an epithet to acquire a different status by exchanging position with other epithets. For example:

(4) Three most important French film directors

As we shall see, nationality should be closer than any other class of epithets to the head, but, depending on the meaning intended, the language user may subvert the rule and produce the following:

(5) A French fast car *vs* an Italian fast car

This means that language users make mental classes of the modifying terms and collocate them in their appropriate classes when communicating. ("Fast car" is here received as a compound lexical unit, where, only superficially, outwardly, "car" is modified by "fast"). This is done automatically by the native speaker of the language concerned, who knows what is appropriate and what is not, and who is automatically (though unconsciously) aware of the implications of position in signifying. The syntactic rules regulating the use of epithets are established by use and are formalised and confirmed by the code. To speak of English premodification, the order of collocation of epithets is regulated by the principle that “the more accidental, subjective and temporary” (Gramley and Patzold 1992: 183-4) is further removed (further to the left hand side of the head) than “the more essential, objective and permanent” (Gramley and Patzold 1992: 183-4)’ which is collocated in graded proximity to the head, the recommended order being: evaluation, size, shape, present participle or past participle, age, colour, nationality or origin.[3] Of course, as far as grammatical rules are concerned those adjectives that can adverbially be modified can also be modified when performing the role of premodifiers.

[3] In the "fast car" example, "fast" is stated as a more permanent feature than nationality.

Structurally, right-hand-side modification (post-modification) has for its main feature its being explicitly declared by linkers, such as prepositions, relative pronouns, or morphological endings, such as *-ING*, *-ED*, or grammatical operators (the preposition *TO* to signal the infinitive form of the verb): be it said in passing that these realisations can always be interpreted as reduced realisations of ellipted relative postmodifying clauses. The full explosion of the relative clause would take place whenever explicitness is required; notably when a new notion (experiment, procedure, etc.) is in order.

A little aside here: it must be pointed out that postmodification should perhaps be described as always being restrictive,[4] and that apposition should perhaps not be listed among the possible postmodifying strategies, in that, first, it provides accessory information, very often information that is already shared by source and target, and occurs between commas; second, and this may be even more relevant, when the appositive unit is spoken the volume of voice and speed of the articulation occur at a level and rhythm other than that of the syntactic chain of the utterance it is said to modify.

It has been noted by various SE (Special English) scholars that the NP constantly undergoes development from postmodified NP to premodified NP to acronym (at least as far as the English language is concerned). Postmodification is used to introduce the new experiment, the new notion, the new procedure. Consolidation in the encyclopaedia leads to premodification. The acronym is the desired linguistic state of the notion when this is frequently used and the components of the notion are numerous. Acronyms respond to the principle of linguistic economy more radically than premodification.

Prepositions and the other linkers can all be identified semantically by a limited number of traits (which in the case of prepositions does not mean at all that the traits are necessarily few and entirely unambiguous: *FOR* may mean destination, “for the airport”; finality, “for a good cause”; concession, “for all the talk done” and so on); however, their various meanings are reasonably controlled and checked by the fact of their complementisers belonging to restricted semantic ranges, and that their realisations occur in specific syntactic patterns.

[4] Against the opinion stated by Quirk and Greenbaum (1990: 364).

If a feature of postmodification is that it can extend so to say endlessly, another is that each complete postmodifying unit is tendentially made up of a limited number of elements: the postmodifying embedded NP is tendentially short, or at least structurally simple, the premodification it embeds is kept to a minimum:

(6) ... discrimination of "*false memories*" in *antispectrum disorder*

(7) … between *appropriateness* of *primary therapy*

(8) … for *early stage carcinoma* and *increased use* of *breast-conserving surgery* (authors' italics)

The last example provides evidence of length of the postmodifying element but confirms its structural elementariness: "increased use of breast-conserving surgery" is a syntagmatic unit, made up of two units, of which the second is postmodifier of the immediate left-hand side constituent.

Postmodification by a relative clause of course uses more words but, as shown by these examples, the relative clause itself is strucurally simple:

(9) Five studies *that were described in four reports* comprised 128 patients with cancer pain

(10) The Honolulu Heart Program is a longitudinal epidemiological study of cardiovascular disease *which began with 8006 Japanese/American men* (authors' italics).

This second example leads to the issue of ambiguous reference in the continuous chain in postmodification.[5] *WHICH*, here, may grammatically refer both to *DISEASE* and to *STUDY*. Readers know that semantically it does not when they have taken in the

[5] An amusing (Italian) example is the information advertised in a shoe-shop window in the Italian Riviera: "*Scarpe per bambini di gomma*" ("rubber children's shoes" or "children's rubber shoes"). The subsequent modifier of a NP with more than one modifying group may sometimes short-circuit with the immediate preceding constituent of the phrase and cause a moment of semantic panic. But well-written texts do not as a rule do so.

completion of the phrase (which comes to an end with "with 8006 Japanese/American men"); they go through a 'bottom-up' procedure which leads to selecting *STUDY* as the appropriate referent because "8006 Japanese/American men" may be *STUDIED* and not *BEGUN*.

But before that step is made possible, they remain open to the fact that the anaphoric referent might well be *DISEASE*; the sentence might, for example, head in this other direction: " ... a longitudinal epidemiological study of a cardiovascular disease *which began with 8006 Japanese/American men becoming affected in one year*".

Ambiguity may also arise frequently when prepositional modifying phrases are used, but as ambiguity is a feature of inaccurate writing we will not expand on the subject.

4. Noun phrase disambiguation

As we noted premodification is characterised by position, because all declarative grammatical links are dropped. As stated above, epithets occupy slots and must be interpreted in terms of the slots occupied and of their relation to each other; since some slots may not be filled, the voids will inevitably be occupied by epithets apparently not in their right position in the syntactic chain; for example nationality may be dropped in *THREE MOST IMPORTANT FILM DIRECTORS*, where *IMPORTANT* becomes the last epithet in the chain before *FILM DIRECTOR*: it must be assumed that *FILM DIRECTOR* is one compound lexical item (*FILM-DIRECTOR*): *FILM* could be deleted without making the term ambiguous only if the phrase were contextualised. Of course *IMPORTANT* here actually precedes nationality slot zero.

As for other aspects of grammar, where scientific texts do not subvert the grammar of the code used but use its potential according to the local needs of scientific discourse, so premodification employs the canonic slots after adjusting them to its own needs. Evaluative and aesthetic epithets are not often required in scientific discourse, while descriptive ones are the order of the day. In *INTENSIVE BLOOD GLUCOSE CONTROL*, all the premodifying words used are quantitative and material. *INTENSIVE*, the one ambiguous adjective, does not provide any *EMOTIONAL* value, but describes a standard modality of action.

Consider the following examples, randomly taken from a variety of medical and technical sources:

(11) autism spectrum disorder

(12) left-ventricular systolic dysfunction [and heart failure]

(13) randomised, placebo controlled cross over study

(14) angiotensin converting enzyme inhibitor (lisinopril)

(15) endocardiographic Heart of England screening study

(16) V-groove tongue- and- groove redwood siding

Disambiguation can only take place if the reader is capable of seeing the corresponding correct semantico-syntactic completion, i.e. can reconstruct what an 'original' postmodification of the head would be. It is assumed that the NP has been identified correctly and its head is the last word in the sequence, in such a way that the examples would be paraphrased as follows:

(11b) spectrum disorder of autism. A systematic description of forms of autism.

(12b) systolic dysfunction of the left ventricle. A dysfunction concerning the contractions of the left ventricle

(13b) cross-over study implying the use of placebos on patients that have been selected at random. A research which has taken place under the control of placebo and was done on patients to whom were administered both medical drugs and placebos and after random selection.

(14b) inhibitor of the enzyme that converts angiotensin.

In the case of example 14 parenthesised *LISINOPRIL* would often be considered an element of the NP chain, as the phrase could be paraphrased as follows: "lisinopril, which is an inhibitor etc. " A problem arises here (akin to the one discussed above on the restrictiveness/non-restrictiveness of apposition) about whether apposition can appropriately be considered as contributing to postmodification, as it is generally inserted in a sentence by means of commas: even more dramatically, here, it is contained within parentheses. As stated above, if any spoken utterance containing apposition is considered, the listener 'hears' that what makes up the apposition belongs in a different phonological chain. One would feel entitled to say it belongs in a different syntactic chain altogether; apposition is perceived as if it were the insertion into a unit of discourse of a fragment coming from another, locally parallel unit of discourse.

(15b) a screening study conducted under the aegis of the Foundation Heart of England on the state of health of the inhabitants of a particular area in the UK, as far as the cardiac organ is concerned.

(16b) [roof] siding made of redwood which has a V-shaped tongue to be fitted in a groove that is also V-shaped (alternatively: siding made of redwood which has a V-shaped groove to be fitted to a tongue that is also V-shaped).

Three to four elements are the vast majority of premodifying realisations in SE (Gotti 1991: 73), but it must be noted that modifying elements often go in pairs, as the examples discussed show (*LEFT-VENTRICULAR, PLACEBO CONTROLLED, CONVERTING ENZYME*). Sometimes the binary item is explicitly indicated by the insertion of a hyphen, as the examples also show: the modification is compound-complex for the sake of simplifying the communication of a text that is complex anyway.

It is a fact that complex premodification will follow complex postmodification, and will be a step further removed from a state of affairs where syntactic complexity is made more by a periphrastic strategy which, however, encodes semantic explicitness; but development from more (postmodification) to less (premodification) syntactic complexity is a syntactic short-cut that can be devised only after an itinerary has become familiar.

In discourse strategies first and following mentions require their own peculiar traits for communication to be successful. Some traits are predicated of an object till those traits become part and parcel with the object. First a book is described as *BORING*, then it becomes *THE BORING BOOK*, first we think of *CLOTHES FOR*

CHILDREN then we word the phrase as *CHILDREN'S CLOTHES*. It is a fact that, when premodification is realised through syntactic reduction (ellipsis of expressed links) and syntactic reorganisation, it is only because the traits now indicated by premodification have been identified and memorised, after a first explicit mental or material statement, as being somewhat inherent in the object described. Interpretation will take place according to procedures that will probably remain unconscious, unperceived, at least till one asks what one meant, or one is asked to write a dictionary entry.

Where the SE variety is concerned, it must be assumed that in the general language user the disambiguating process will take place in more time and will be more laborious than in the expert; time and method of realisation will require some different degree of conscious application of some paraphrasing process, which, however, is unlikely to be substantially different between same language users. The mind of the language user cannot not be under a measure of strain. The contention here is that all language users are under that strain, the native speaker and the expert as well as the language learner and the translator. The difference lies only in the different paths followed, whether that path is the road on the map or a shortcut.

5. Conclusion

It is suggested that NP disambiguation strategies are applied at the grammatical, syntactic and semantic levels in this order. The grammatical and syntactical levels are the surface levels, they are embodied by the explicit traits of pre- and post-modification, where prepositions, relative pronouns and morphological markers – grammar – or their absence, and sequence, which replicates either the world observed or the observer's point of observation – syntax – are declared in discourse. The semantic level is activated whenever ambiguities arise. Sometimes more than semantic disambiguation is required, as in the case of *V-GROOVE TONGUE-AND-GROOVE REDWOOD SIDING*, where only ‘knowledge of the world’ will be decisive in making the right visualisation.

The complexity of NP modification is such, however, that the present paper does not claim to having dealt with the problem exhaustively. Its only aim is to offer points of reflection for the examination of wider ranges of texts and text types leading eventually to the definition of rules presiding over the English NP syntactic construction.

References

Bolinger, D. (1977) *Meaning and Form*, Longman, London.

Gotti, M. (1991) *I linguaggi specialistici*, La Nuova Italia, Firenze.

Gramley, S. and K.M. Patzold (1992) *A Survey of Modern English*, Routledge, London.

Greenbaum, S. and R. Quirk (1990) *A Student's Grammar of the English Language*, Longman, London.

Halliday, M.A.K. [1985] (1994) *An Introduction to Functional Grammar*, Arnold, London.

Quirk, R. and S Greenbaum (1973) *A University Grammar of English*, Longman, London.

Quirk, R., S. Greenbaum, G. Leech, G. and J. Svartvik (1985) *A Comprehensive Grammar of the English Language*, Longman, London.

Sager, J., D. Dungworth and P.F. McDonald (1980) *English Special Languages*, Oscar Brandstetter Verlag, Wiesbaden.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 153-169 © 2008 Polimetrica International Scientific Publisher Monza/Italy

What dictionaries leave out: new non-adapted Anglicisms in Italian

Cristiano Furiassi – University of Turin

1. Introduction

In a vast repertoire of neological creativity, English words represent the most productive storehouse from which the Italian language draws everyday. New non-adapted Anglicisms frequently appear in different types of texts, especially in newspapers (Fanfani 2003), although many of them are not destined to survive.

The phenomenon is evident and has been recently exploited by some major Italian dictionary editors and publishers in order to increase the number of non-adapted Anglicisms in their wordlists (Furiassi, forthcoming). However, if on the one hand, several non-adapted Anglicisms included in dictionaries seem superfluous or more pertinent to specialized or *ad hoc* dictionaries, on the other hand, one may find that some non-adapted Anglicisms heard or read by the Italian speaker are surprisingly not recorded. Indeed, according to De Mauro:

> Ogni tanto incontriamo nei giornali, in tv, nei discorsi, una parola che ci pare nuova. Se abbiamo sotto mano un buon dizionario, nella maggior parte dei casi ci accorgiamo che la parola è già là, bella (o brutta) e registrata. Ma altrettanto spesso circolano nell'uso parole non ancora registrate dai dizionari, anche i più accurati e i più ampi. (De Mauro 2006: v)

The present work is aimed at detecting new non-adapted Anglicisms which are not included in recent Italian monolingual dictionaries, i.e. De Mauro (2000, 2003), Devoto and Oli (2006), Sabatini and Coletti

(2005), and Zingarelli (2006), but which seem familiar enough to qualify for inclusion in prospective updating.

The search for non-adapted Anglicisms was carried out on a large corpus of Italian newspaper language, i.e. La Repubblica corpus, by means of n-gram based queries. Since the La Repubblica corpus covers a time frame of sixteen years – between 1985 and 2000 – whereas the dictionaries consulted were all published between 2003 and 2006, the time span considered should have been reasonably sufficient for editors and publishers to detect and include non-adapted Anglicisms in their dictionaries.

In the initial part of the article n-grams are described with special focus on how they provide scientific evidence of grapheme typicality for English and Italian respectively. Then, an account is given of the procedures applied in the automatic search for n-gram based lists of English words in the La Repubblica corpus. In addition, the semi-automatic refinement of the lists obtained from the corpus through an open-source word-processing tool, i.e. OpenOffice Writer, is described. The manual refinement of the lists obtained is also explained. Finally, the non-adapted Anglicisms found by comparing the lists extracted from the corpus with an existing wordlist of non-adapted Anglicisms taken from a dictionary of foreign words in Italian, i.e. De Mauro and Mancini (2003), are cross-checked against the wordlists of Italian monolingual dictionaries in order to be analyzed and discussed.

2. N-grams: scientific evidence of grapheme typicality

It is often taken for granted that certain graphemes or grapheme sequences are more likely to appear in one language rather than another. N-grams, which provide the scientific evidence needed to attest grapheme typicality, are defined as follows:

> *N*-grams are recurrent combinations of items of various kinds (e.g. graphemes, morphemes, words, phrases, and sentences) which may be customised according to the user's needs. *N*-grams are useful for several linguistic functions – including finding collocations for machine translation and automatic tagging of texts – and provide insights into language usage. (Furiassi and Hofland 2007: 355)

Table 1 below, which is an extract from Furiassi and Hofland (2007: 356-357), shows the 16 n-grams most typical of English.[1]

n-gram sequence	**n-gram frequency (English)**	**n-gram frequency (Italian)**	**typicality (%)**
-Y	4,442	1	99.98
-NG	4,353	2	99.95
-HO-	1,636	2	99.88
W-	1,511	2	99.87
-ES	3,071	5	99.84
-ED	6,021	14	99.77
H-	2,022	5	99.75
-HA-	1,627	6	99.63
-S	12,245	52	99.58
-D	7,096	54	99.24
-G	4,463	37	99.18
-TION-	2,045	31	98.51
-TIO-	2,097	38	98.22
-N	3,374	101	96.32
-T	3,103	200	93.94
-R	2,545	853	74.90

Table 1. The 16 n-grams most typical of English.

The 16 n-grams in Table 1 were chosen since they are much more typical of English, e.g. there is a 99.98 % chance that a word ending with *-Y* belongs to English rather than Italian.[2] The first column shows each single n-gram in order of typicality. The second and third columns show how typical each n-gram is for English and

[1] The procedures employed to produce n-gram statistics are described in Furiassi and Hofland (2007: 357).

[2] Although the n-gram *-R* is 'only' 74.90 % more typical of English than Italian, it was considered for the analysis since the frequency of the grapheme *-R* in final position is biased by the high frequency of the preposition *PER* in the Italian language.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

Italian respectively.[3] The fourth column shows the degree of typicality as a percentage.[4]

The hyphen indicates beginning of word or end of word, whether it follows or precedes the string of graphemes. For instance, words beginning with *W*- are 99.87 % more typical of English than Italian. Strings with hyphens on both sides show that a certain sequence of graphemes is found within a certain word. For example, words including -*HO*- are 99.88 % more typical of English than Italian.

In the analysis, n-grams of graphemes were used to recognize the most typical patterns of English and Italian respectively and to identify English words in Italian, i.e. non-adapted Anglicisms, according to the probabilities that certain graphemes or grapheme combinations have to appear in either of the two languages.

3. The search for non-adapted Anglicisms

N-gram based wordlists containing non-adapted Anglicisms were searched for in the La Repubblica corpus. The size of the La Repubblica corpus amounts to 380,887,318 tokens. As indicated by Baroni *et al.* (2004) and Aston and Piccioni (2004), the corpus includes 593,593 articles that appeared in 5,163 issues published between 1985 and 2000 by the national daily La Repubblica – the second most widely-read Italian newspaper. The corpus is also tokenized, POS-tagged, lemmatized, and categorized in terms of genre and topic.

In order to search for n-gram based wordlists in the La Repubblica corpus, regular expressions were typed in the corpus query processor (CQP) provided (Evert 2005: 29-31).[5] The CQP-style regular expressions used in the queries are as follows:

[3] The frequency figures included in the second and third columns represent the raw occurrences of n-grams provided by the procedures described in Furiassi and Hofland (2007).

[4] The degree of typicality is calculated by dividing the English n-gram frequency by the sum of the English n-gram frequency and the Italian n-gram frequency: the result is then represented as a percentage. For example, the degree of typicality of the n-gram *H*- is computed as follows: 2,022 / (2,022 + 5) = 0.9975 = 99.75 %.

[5] The author is particularly grateful to Guy Aston and Marco Baroni – *Scuola Superiore di Lingue Moderne per Interpreti e Traduttori* (*SSLMIT*) and *Università*

1. word+0=/.*s/ pos+0=/NOUN/
2. word+0=/.*d/ pos+0=/NOUN/
3. word+0=/.*g/ pos+0=/NOUN/
4. word+0=/.*y/ pos+0=/NOUN/
5. word+0=/.*n/ pos+0=/NOUN/
6. word+0=/.*t/ pos+0=/NOUN/
7. word+0=/.*r/ pos+0=/NOUN/

The above regular expressions (1-7) generated frequency lists of nouns ending with *-S*, *-D*, *-G*, *-Y*, *-N*, *-T*, and *-R* respectively.

8. word+0=/h.*/pos+0=/NOUN/
9. word+0=/w.*/ pos+0=/NOUN/

The above regular expressions (8,9) generated frequency lists of nouns beginning with *H-* and *W-* respectively.

10. word+0=/.*tio.*/ pos+0=/NOUN/
11. word+0=/.*ho.*/ pos+0=/NOUN/
12. word+0=/.*ha.*/ pos+0=/NOUN/

The above regular expressions (10-12) generated frequency lists of nouns including *-TIO-*, *-HO-*, and *-HA-* respectively.

Only 12 n-gram based wordlists out of 16 were requested: the reason for this is that the search for *-TIO-* already included items with *-TION-*, the search for *-S* included items ending with *-ES*, the search for *-D* included items ending with *-ED*, and the search for *-G* included items ending with *-NG*. In addition, n-grams were searched for within the word class of nouns only, since almost all non-adapted Anglicisms in Italian are nouns (Pulcini 2002: 159-161).

4. Refinement procedures

The queries carried out generated 12 lists containing n-gram based items which included English words. Table 2 shows the figures representing how many items were included in each list.

degli Studi di Bologna – for allowing on-line access to the corpus and providing technical support.

n-gram based wordlist	items provided by the CQP queries	items with a raw frequency above or equal to 4	items after all refinement procedures including duplicates	items after all refinement procedures excluding duplicates
-S	13,111	3,089	1,455	1,144
-R	6,030	1,443	775	527
-T	5,815	1,211	576	505
-N	4,192	1,051	492	407
-G	2,674	731	610	471
-HA-	2,513	532	165	118
-HO-	2,206	448	252	167
W-	1,951	303	183	153
-Y	1,928	509	384	345
-TIO-	1,884	422	227	192
H-	1,883	425	201	157
-D	1,396	320	196	154
total	45,556	10,484	5,516	4,340

Table 2. The 12 n-gram based wordlists including non-adapted Anglicisms.

The table is arranged according to the number of items included in each list after automatic extraction. The first column shows each n-gram used to identify lists of items including non-adapted Anglicisms in the corpus; the second column shows the figures for each list considering all items provided by the CQP queries, for a total of 45,556 items (see *Automatic refinement*); the third column shows the figures for each list considering items with a raw frequency above or equal to 4, i.e. more than 1 per 100 million words, for a total of 10,484 items (see *Automatic refinement* and *Semi-automatic refinement*); the fourth column shows the figures for each list after the refinement procedures including duplicates, for a total of 5,516 items (see *Manual refinement*); the fifth column shows the figures for each list after the refinement procedures excluding duplicates, for a total of 4,340 items (see *Manual refinement*).

Automatic refinement

The 45,556 items included in the 12 lists extracted from the La Repubblica corpus (see Table 2, second column) were too many to be analyzed manually. Therefore, in order to exclude *hapax legomena* and low frequency entries, only items with a raw frequency above or equal to 4, i.e. with an absolute frequency above 1 per 100 million words, were considered. Since the lists were provided in order of decreasing frequency, this procedure was carried out automatically by simply deleting the items at the end of each list with a raw frequency between 1 and 3. This first step reduced the number of items included in the 12 lists to 10,484 (see Table 2, third column).

Semi-automatic refinement

As expected, besides non-adapted Anglicisms, misspellings, hybrids, and several words from languages other than English were also present in the lists. However, manually checking the 10,484 items included in the 12 automatically-refined lists would still have required an excessive amount of time. Therefore, a semi-automatic refinement of the lists was applied.

With the aid of the built-in dictionary and the spell-checker of the word-processor OpenOffice Writer, spelling mistakes made by the authors of the articles, e.g. *BREAFING* or *BREEFING vs BRIEFING*, hybrids,[6] e.g. *ACQUAPLANING vs AQUAPLANING*, and non-English words, e.g. *METROPOLIS*, *MYTHOS* (Greek), *CARITAS*, *MAGISTER*, *MODUS*, *OPUS* (Latin), *DESPERADOS*, *TOREADOR* (Spanish), *CARNETS*, *REVENANTS* (French), were eliminated. All these items were easily retrievable since they were underlined by the spell-checker provided by OpenOffice Writer.[7] However, the semi-automatic

[6] Indeed several Italian-English hybrids exist in the Italian language. Among the most productive combining forms which appeared in the lists, the following were found: *-BABY*, e.g. *PENSIONE-BABY*, *-BOY*, e.g. *MAFIA-BOY*, *-RECORD*, e.g. *TEMPO-RECORD*, *-MANAGER*, e.g. *MEDICO-MANAGER*, and *-KILLER*, e.g. *VINO-KILLER*.

[7] The open-source software OpenOffice Writer identifies whether a certain item belongs to English or not according to the built-in dictionary stored in the software memory. However, items underlined by the spell-checker – set on British English and American English – were not all non-English words, e.g. *STREAKER*. This is due to the fact that some items were not included in the built-in dictionary.

refinement did not allow the researcher to eliminate proper names, acronyms and abbreviations, and items including '@' and 'www'. This refinement procedure had to be carried out manually.

Manual refinement

The procedures used in manually reducing the lists after the semi-automatic refinement were carried out in the following order: proper names were excluded by deleting words beginning with a capital letter, e.g. *APARTHEID*; capitalized acronyms and abbreviations were excluded by deleting sequences of capital letters, e.g. *ABS*; and items including @ and *WWW* were excluded since they were part of either e-mail addresses or web-sites.

After these procedures were carried out, the 12 lists were merged into one list which contained 5,516 items (see Table 2, fourth column). The list was finally rearranged in alphabetical order to make the manual exclusion of duplicates easier.[8] The label 'duplicates' is here employed in a broad sense since it not only includes items that appeared in more than one list, but also comprises hyphenated compounds, e.g. *WEEKEND vs WEEK-END*, alternative British English and American English spelling variants, e.g. *HUMOUR vs HUMOR*, and plurals, e.g. *HAMBURGER vs HAMBURGERS*.[9]

After duplicates were eliminated, a final merged list of 4,340 items was obtained (see Table 2, fifth column). This merged list underwent further manual refinement consisting of the exclusion of eponyms, e.g. *CARTER*, ethnonyms, e.g. *AMERICANS*, *ITALIANS*, toponyms, e.g. *WATERGATE*, trademarks,[10] e.g. *DOLBY®*, *PING PONG®*, *WALKMAN®*, *WONDERBRA®*, units of measurement,[11] e.g.

[8] For instance, as regards duplicates, some items found by searching for *-D* were the same items found by searching for *H-*, e.g. *HEAD*.

[9] Plurals were eliminated throughout whenever the singular form was also attested. For instance, both *AIRBAG* and *AIRBAGS* were counted in the merged list which included duplicates (see Table 2, fourth column) but only *AIRBAG* was kept in the final merged list which excluded duplicates (see Table 2, fifth column). The plural form was kept only when the singular was absent, e.g. *HOUSES*.

[10] De Mauro and Mancini (2003: 345-346) list trademarks as "nomi commerciali" in a separate section. For a comprehensive exclusion of trademarks see Furiassi (2006: 204-206).

[11] De Mauro and Mancini (2003: 946) list units of measurement as "unità di misura" and toponyms as "nomi geografici" in a separate section.

BRENT, non-capitalized acronyms, e.g. *ABS*, and all items that the spell-checker considered English in spite of the fact that they were borrowed from other languages, e.g. *AMATEUR* (French), *CONDOR* (Spanish), *DELICATESSEN* (German), *GINSENG* (Chinese), *HABITAT* (Latin), *HAREM* (Arabic), *CHAOS* (Greek), *ROBOT* (Czech). False Anglicisms,[12] e.g. *FOOTING*, and all items including prefixes of Latin or Greek origin were also excluded,[13] e.g. *GEO-MARKETING*. Finally, English grammatical items, e.g. *BEEN*, *COULD*, *FOR*, *IT*, *WHAT*, which were part of titles of songs, books, and movies, e.g. *"THAT'S WHAT FRIENDS ARE FOR"*, and were therefore included in the search for NOUN in the initial CQP queries, were eliminated. This final list amounted to 2,816 items.

5. Comparing corpus-based with dictionary wordlists

The refined list including 2,816 non-adapted Anglicisms, which was obtained after the automatic, semi-automatic, and manual refinement procedures described above, was checked against the wordlist of 5,510 non-adapted Anglicisms in Italian recorded by De Mauro and Mancini (2003: 905-932).[14] Before comparing the final merged list with each Italian monolingual dictionary considered, De Mauro and Mancini (2003) was used as the exclusion corpus in order to reduce further the number of non-adapted Anglicisms extracted from the La Repubblica corpus.

[12] De Mauro and Mancini (2003: 932) list false Anglicisms as "pseudoinglese" separately. For a comprehensive exclusion of false Anglicisms see Furiassi (2003: 128-133).

[13] The complete list of combining forms of Latin or Greek origin is as follows: *AGRI-*, e.g. *AGRI-BUSINESS*, *ANTI-*, e.g. *ANTI-STRESS*, *AUTO-*, e.g. *AUTO-TEST*, *BIO-*, e.g. *BIO-ENGINEERING*, *CO-*, e.g. *CO-LEADER*, *DE-*, e.g. *DE-ESCALATION*, *ECO-*, e.g. *ECO-AUDIT*, *EURO-*, e.g. *EURO-BUSINESS*, *EX-*, e.g. *EX-BOSS*, *GEO-*, e.g. *GEO-MARKETING*, *MAXI-*, e.g. *MAXI-YACHT*, *MEGA-*, e.g. *MEGA-HOLDING*, *MICRO-*, e.g. *MICRO-COMPUTER*, *MINI-*, e.g. *MINI-DERBY*, *MULTI-*, e.g. *MULTI-TASKING*, *NEO-*, e.g. *NEO-MANAGER*, *PORNO-*, e.g. *PORNO-SHOW*, *POST-*, e.g. *POST-HIPPIES*, *PRE-*, e.g. *PRE-SELECTION*, *RE-*, e.g. *RE-STYLING*, *RETRO-*, e.g. *RETRO-FIT*, *SUB-*, e.g. *SUB-HOLDING*, *SUPER-*, e.g. *SUPER-GANGSTER*, *TELE-*, e.g. *TELE-STAR*, *VICE-*, e.g. *VICE-CHAIRMAN*, *VIA-*, e.g. *VIA-COMPUTER*, and *VIDEO-*, e.g. *VIDEO-SHOP*.

[14] De Mauro and Mancini (2003) is a dictionary which contains all non-adapted borrowings in Italian – including non-adapted Anglicisms – which are part of De Mauro (2000) and De Mauro (2003), a 7-volume Italian monolingual dictionary which includes 251,209 main entries.

Unexpectedly, as many as 1,366 non-adapted Anglicisms present in the merged list were absent from De Mauro and Mancini (2003). Such a high number may be justified by the fact that the merged list still included English lexical items other than nouns, i.e. adjectives, lexical verbs, and adverbs, which were part of titles of songs, books, and movies, e.g. *"FRIED GREEN TOMATOES AT THE WHISTLE STOP CAFÉ"*, and were therefore included in the search for NOUN in the initial CQP queries.

6. New non-adapted Anglicisms

Given the surprisingly large number of non-adapted Anglicisms which did not match those included in De Mauro and Mancini (2003), for the purpose of the present work a selection of the most interestingly familiar non-adapted Anglicisms was made.

The 35 non-adapted Anglicisms selected are *BANDLEADER*, *BOOK SHOP*, *CLASS ACTION*, *COACHING*, *COFFEE-SHOP*, *CREATIVE WRITING*, *CRIME STORY*, *DECISION MAKING*, *FIDELITY CARD*, *GALLERY*, *GOVERNMENT*, *HIT-MAN*, *INKJET*, *LIBRARY*, *MISTRESS*, *POLICY*, *RAVE PARTY*, *RECRUITING*, *REGULAR SEASON*, *RIDER*, *ROCKBAND*, *SERVE AND VOLLEY*, *SETTING*, *SHOOT OUT*, *SHOW-WOMAN*, *SOLD OUT*, *SONGWRITER*, *SPLIT SCREEN*, *SPONSORING*, *STAGFLATION*, *STREAKER,* *TEAM MANAGER*, *WINE-MAKER*, *WORKAHOLIC*, and *WORKOUT*.

Each item in the above list was checked in the recent editions of some Italian monolingual dictionaries, i.e. Devoto and Oli (2006), Sabatini and Coletti (2005), and Zingarelli (2006). Among the 35 non-adapted Anglicisms listed, Devoto and Oli (2006) record 9 entries: *BOOK SHOP*, *CLASS ACTION*, *CREATIVE WRITING*, *FIDELITY CARD*, *RECRUITING*, *RIDER*, *SETTING*, *SOLD OUT*, and *WORKAHOLIC*; Sabatini and Coletti (2005) record 9 entries: *BOOK SHOP*, *COACHING*, *CREATIVE WRITING*, *RAVE PARTY*, *REGULAR SEASON*, *RIDER*, *SHOW-WOMAN*, *SOLD OUT*, and *STAGFLATION*; and Zingarelli (2006) records 2 entries: *FIDELITY CARD* and *RIDER*. Also *TEAM MANAGER* was found as a sub-entry in Sabatini and Coletti (2005). Therefore 15 items, i.e. *BOOK SHOP*, *CLASS ACTION*, *COACHING*, *CREATIVE WRITING*, *FIDELITY CARD*, *RAVE PARTY*, *RECRUITING*, *REGULAR SEASON*, *RIDER*, *SETTING*, *SHOW-WOMAN*, *SOLD OUT*, *STAGFLATION*, *TEAM MANAGER*, and *WORKAHOLIC* were excluded from the list of the 35 non-adapted Anglicisms selected.

Non-adapted Anglicisms which were not found in the Italian monolingual dictionaries considered were also checked in Adamo and Della Valle (2003: 3-57) and Adamo and Della Valle (2005: 5-15) respectively.[15] In Adamo and Della Valle (2003), besides 6 non-adapted Anglicisms which were already included in the wordlists of the dictionaries considered, i.e. *BOOK SHOP*, *RECRUITING*, *SHOW-WOMAN*, *SOLD OUT*, and *WORKAHOLIC*, only *CRIME STORY* was found. In Adamo and Della Valle (2005), on the other hand, besides 1 non-adapted Anglicism which was already included in the wordlists of the dictionaries considered, i.e. *CLASS ACTION*, only *WINE-MAKER* was found. The comparison between the list of non-adapted Anglicisms selected and Adamo and Della Valle (2003) and Adamo and Della Valle (2005) is acceptable since Adamo and Della Valle (2003: x; 2005: vii) state that they have used as 'corpus di esclusione', among other sources, De Mauro (2000) and De Mauro (2003).[16] Therefore, 2 more items, i.e. *CRIME STORY* and *WINE-MAKER*, were excluded from the list including the 35 non-adapted Anglicisms selected.

A final selection of 18 new non-adapted Anglicisms which were not found in any of the dictionaries consulted, including dictionaries of neologisms and foreign words, is presented in Table 3 below.

[15] Adamo and Della Valle (2003) and Adamo and Della Valle (2005) are two dictionaries which collect neologisms recently encountered in the Italian press.

[16] It is important to consider that, although *DECISION MAKING* was not found in any of the dictionaries consulted, *DECISION-MAKER* is recorded by Sabatini and Coletti (2005). In addition, even though *STREAKER* was never found, *STREAKING* is recorded in Devoto and Oli (2006), Sabatini and Coletti (2005), and Zingarelli (2006). Also *SONGWRITER* was never found in the dictionaries consulted, although *SONGWRITING* appears in Adamo and Della Valle (2003). Consequently, the non-adapted Anglicisms *DECISION MAKING*, *SONGWRITER*, and *STREAKER* were included in Table 3.

Anglicism	two-word compound	one-word compound	hyphenated compound	total
ROCKBAND	70/0	119/0	10/0	199
SERVE AND VOLLEY	181/0	-	5/0	186
LIBRARY	-	69/3	-	72
BANDLEADER	10/0	26/0	5/0	41
SONGWRITER	1/0	35/1	4/0	41
COFFEE-SHOP	9/2	0/0	26/1	38
INKJET	15/0	5/0	15/0	35
SHOOT OUT	9/1	0/0	8/0	18
POLICY	-	13/3	-	16
DECISION MAKING	8	0	5	13
SPONSORING	-	13	-	13
HIT-MAN	0/3	0/1	0/5	9
SPLIT SCREEN	5/0	0/0	4/0	9
GOVERNMENT	-	7/0	-	7
WORKOUT	0/0	6/0	1/0	7
MISTRESS	-	4/1	-	5
STREAKER	-	5/0	-	5
GALLERY	-	0/0	-	0

Table 3. A selection of 18 new non-adapted Anglicisms.

The first column lists the 18 new non-adapted Anglicisms selected in raw-frequency order; the second column displays the frequency scores for two-word compounds; the third column displays the frequency scores for one-word compounds; the fourth column displays the frequency scores for hyphenated compounds; the fifth column displays overall frequency scores. Slashes are used to separate the frequency scores for singular and plural forms. For instance, the item *SHOOT OUT*, as a two-word compound, was found 9 times in the singular form and once in the plural, i.e. *SHOOT OUTS*, and, as a hyphenated compound, 8 times in the singular form, i.e. *SHOOT-OUT*.

Both singular and plural forms, except for *DECISION MAKING* and *SPONSORING*, and all compound variants were counted in the La Repubblica corpus in order to provide reliable frequency scores. For instance, each concordance line was manually checked to exclude from the counts instances of non-adapted Anglicisms belonging to proper nouns, e.g. *BODLEIAN LIBRARY*.

Among the 18 non-adapted Anglicisms which are part of the list in Table 3, 6, i.e. *BANDLEADER*, *COFFEE-SHOP*, *LIBRARY*, *ROCKBAND*, *SERVE AND VOLLEY*, and *SONGWRITER*, have a raw frequency between 38 and 199, which is roughly between 10 and 50 per 100 million words; 11, i.e. *DECISION MAKING*, *GOVERNMENT*, *HIT-MAN*, *INKJET*, *MISTRESS*, *POLICY*, *SHOOT OUT*, *SPLIT SCREEN*, *SPONSORING*, *STREAKER*, and *WORKOUT*, have a raw frequency between 5 and 35, which is roughly between 1 and 9 per 100 million words; and 1, i.e. *GALLERY*, was never found except in compounds, e.g. *ART GALLERY*, or as part of proper nouns, e.g. *TATE GALLERY*.

Frequency considerations must be taken into account since they provide interesting insight into how often each new non-adapted Anglicism is actually used in the corpus analyzed. As a consequence, corpus-based frequency counts should help lexicographers decide which new non-adapted Anglicisms should be added, with good reason, to the wordlists of future dictionaries (Furiassi, forthcoming).

Although it is difficult to establish a clear-cut threshold frequency on the basis of which new non-adapted Anglicisms should be included in dictionaries (Furiassi, forthcoming), the 6 items, i.e. *BANDLEADER*, *COFFEE-SHOP*, *LIBRARY*, *ROCKBAND*, *SERVE AND VOLLEY*, and *SONGWRITER*, which score above 10 per 100 million words, are the ones that the lexicographer may consider including in general-purpose monolingual dictionaries. The non-adapted Anglicisms which score below 10 per 100 million words might nonetheless be included in dictionaries of Anglicisms. However, well-balanced frequency considerations are of the utmost lexicographic importance especially if "the main focus of the dictionary is not 'exhaustiveness', but 'representativeness'" (Pulcini, forthcoming).

After these quantitative considerations, some qualitative observations are necessary. Some non-adapted Anglicisms coexist with their Italian equivalents. For instance, the non-adapted Anglicisms *INKJET* and *SPONSORING* have an Italian equivalent, which is a calque from English, in the case of *GETTO D'INCHIOSTRO*, or an adapted Anglicism, in the case of *SPONSORIZZAZIONE*. It is worth noticing that the Italian equivalent *SPONSORIZZAZIONE* outnumbers *SPONSORING* whereas *GETTO D'INCHIOSTRO* and *GETTO DI INCHIOSTRO* are not as frequent as *INKJET*.[17]

In regard to the semantic fields of new non-adapted Anglicisms, among the 6 which have a considerable frequency, *BANDLEADER*, *ROCKBAND*, and *SONGWRITER* belong to music, *SERVE AND VOLLEY* to sport, and *COFFEE-SHOP* and *LIBRARY* to other domains.

7. Drawbacks

The main deficit of this study is the fact that, despite the time-consuming refinement procedures carried out, a sufficiently restricted number of non-adapted Anglicisms could not be reached. This generated the startlingly large gap between the refined corpus-based wordlist and the wordlist of De Mauro and Mancini (2003), thus forcing the author to select a limited number of non-adapted Anglicisms.

Although the initial aim of this research was that of finding all the new non-adapted Anglicisms present in the Italian press which are not included in recent editions of some Italian monolingual dictionaries, the selection made from the merged list did nevertheless prove sufficient to provide evidence of the constant input of non-adapted Anglicisms in Italian.

A further drawback lies in having considered a limited number of n-grams at the initial stage of the analysis: more non-adapted Anglicisms could have been found if the list of n-grams, on which the research is based, had been extended, for instance, by including items ending with *-K* and *-L*. Another limitation is that only one newspaper is definitely not representative enough to study the phenomenon thoroughly.

An additional weak point is that the search for non-adapted Anglicisms in the La Repubblica corpus was carried out by considering graphic forms only: differences in meaning between how non-adapted Anglicisms are used in the La Repubblica corpus

[17] The La Repubblica corpus displayed 24 occurrences of *GETTO D'INCHIOSTRO* and 7 of *GETTO DI INCHIOSTRO* against 35 occurrences of *INKJET*, considering both singular and plural forms. The occurrences of *SPONSORIZZAZIONE* were 518 against 13 of *SPONSORING*.

and how they are defined by the dictionaries consulted were not taken into account, e.g. *COACHING*.[18]

Finally, some non-adapted Anglicisms, which were considered new in this research, might be found in the latest editions of the Italian monolingual dictionaries considered. However, it was not possible to consult the most up-to-date editions of these dictionaries since they were all published between spring and summer 2007.

8. Conclusion

Even though quantitative considerations might not be completely watertight due to the drawbacks described, 18 new non-adapted Anglicisms were detected. Therefore, the native speaker's impression that some non-adapted Anglicisms which appear to be current in the press, but are absent from dictionaries, seems to be confirmed. However, Italian monolingual dictionaries, at least those considered, seem quite complete and up to date – as far as the inclusion of non-adapted Anglicisms is concerned – and close to providing a trustworthy representation of language in use.

Certainly 'the bigger the better' is not always a merit in lexicographic practice (De Mauro 2005: 172). However, new non-adapted Anglicisms, which can be considered of broad circulation in various domains and which are excluded from the wordlists of Italian monolingual dictionaries, should definitely be added to future dictionaries. This is particularly true in the case of a dictionary of Anglicisms in Italian (Pulcini 2006: 319-321).

With reference to the methodology applied in the analysis, although the present article does not aim to outline what has been defined by Alex (2005: 133) as "an automatic classifier of foreign

[18] Although excluded from the list of new non-adapted Anglicisms in Italian since it was found in Sabatini and Coletti (2005), according to De Mauro (2006: 19,20), *COACHING* may refer to "allenamento sportivo", "tutoraggio", or – within a team in a company – to an activity which makes employees "sviluppare la propria personalità". However, the definition provided by Sabatini and Coletti (2005) only considers *COACHING* as "periodo di formazione di un direttore d'azienda sotto la guida di un istruttore più esperto". Furthermore, from the concordance lines extracted from the La Repubblica corpus, *COACHING* is used with the specialized meaning of "attività proibita durante le partite ufficiali di tennis consistente nella direzione degli atleti impegnati in gara".

inclusions" properly, it attempted to show how off-the-shelf and daily used word-processing tools combined with n-gram statistics and corpus queries may prove valuable for lexicologists and lexicographers in identifying non-adapted Anglicisms in Italian texts as well as borrowings in other languages.

Despite the fact that further and more accurate quantitative research is needed, hopefully the 'homemade' set of procedures described in this article will allow linguists to investigate the use of non-adapted Anglicisms not only in Italian but in any given language and/or language domain.

References

Adamo, G. and V. Della Valle (2003) *Neologismi quotidiani. Un dizionario a cavallo del millennio: 1998-2003*, Olschki, Firenze.

Adamo, G. and V. Della Valle (2005) *2006 parole nuove. Un dizionario di neologismi dai giornali*, Sperling & Kupfer, Milano.

Alex, B. (2005) "An unsupervised system for identifying English inclusions in German text", in *Proceedings of the Association for Computational Linguistics (ACL) Student Research Workshop*, Ann Arbor, Michigan, pp. 133-138. http://www.cogsci.ed.ac.uk/~balex/PUBLICATIONS/ACL05.pdf

Aston, G. and L. Piccioni (2004) "Un grande corpus di italiano giornalistico", in G. Bernini, G. Ferrari and M. Pavesi (eds), *Atti del III congresso di studi dell'Associazione Italiana di Linguistica Applicata (AItLA)*, Guerra Edizioni, Perugia, pp. 289-310.

Baroni, M., S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston and M. Mazzoleni (2004) "Introducing the La Repubblica corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper Italian", *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Vol. 5, ELRA, Lisbon, pp. 1771-1774.
http://sslmitdev-online.sslmit.unibo.it/corpora/downloads/rep_lrec_ 2004.pdf

De Mauro, T. (2005) *La fabbrica delle parole*, UTET, Torino.

De Mauro, T. (2006) *Dizionarietto di parole del futuro*, Laterza, Roma/Bari.

De Mauro, T. (ed.) (2000) *Grande dizionario italiano dell'uso*, UTET, Torino.

De Mauro, T. (ed.) (2003) *Nuove parole italiane dell'uso del Grande dizionario italiano dell'uso*, UTET, Torino.

De Mauro, T. and M. Mancini [2001] (2003) *Parole straniere nella lingua italiana*, Garzanti Linguistica/UTET, Milano.

Devoto, G. and G.C. Oli (eds) (2006) *Il Devoto-Oli 2007. Vocabolario della lingua italiana*, Le Monnier, Firenze / bSmart, Cesano Maderno.

Evert, S. (2005) *The CQP Query Language Tutorial,* IMS Stuttgart, Stuttgart. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/cqp-tutorial.2up.pdf

Fanfani, M. (2003) “Per un repertorio di anglicismi in italiano”, in A.-V. Sullam Calimani (ed.), *Italiano e inglese a confronto*, Franco Cesati, Firenze, pp. 151-176.

Furiassi, C. (2003) “False Anglicisms in Italian monolingual dictionaries: a case study of some electronic editions”, *International Journal of Lexicography* 16 (2), pp. 121-142.

Furiassi, C. (2006) “Translating American and British trademarks into Italian. Are bilingual dictionaries an aid to the user?”, in F. San Vicente (ed.), *Lessicografia bilingue e traduzione: metodi, strumenti, approcci attuali*, Polimetrica, Monza, pp. 199-213.

Furiassi, C. (forthcoming) “Non-adapted Anglicisms in Italian: attitudes, frequency counts, and lexicographic implications”, in R. Fischer and H. Pulaczewska (eds), *Anglicisms in Europe: Linguistic Diversity in a Global Context*, Cambridge Scholars Publishing, Newcastle-upon-Tyne.

Furiassi, C. and K. Hofland (2007) “The retrieval of false Anglicisms in newspaper texts”, in R. Facchinetti (ed.), *Corpus Linguistics 25 Years On*, Rodopi, Amsterdam/New York, pp. 347-363.

La Repubblica: www.repubblica.it

La Repubblica Corpus: sslmit.unibo.it/repubblica

OpenOffice: www.openoffice.org

Pulcini, V. (2002) “Italian”, in M. Görlach (ed.), *English in Europe*, Oxford University Press, Oxford, pp. 151-167.

Pulcini, V. (2006) “A *New Dictionary of Italian Anglicisms*: the aid of corpora”, in E. Corino, C. Marello and C. Onesti (eds), *Proceedings XII EURALEX International Congress*, Vol. 1, Edizioni dell’Orso, Alessandria, pp. 313-322.

Pulcini, V. (forthcoming) “*A Dictionary of Italian Anglicisms*: criteria of inclusion and exclusion”, in G. Iamartino and N. Brownlees (eds), *Insights into English and Germanic Lexicology and Lexicography: Past and Present Perspectives*, Polimetrica, Monza.

Sabatini F. and V. Coletti (eds) (2005) *Il Sabatini-Coletti 2006. Dizionario della lingua italiana*, Rizzoli/Larousse, Milano / Expert System, Modena.

Zingarelli, N. (ed.) (2006) *Lo Zingarelli 2007. Vocabolario della lingua italiana*, Zanichelli, Bologna / I.CO.GE Informatica, Trento.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 171-187 © 2008 Polimetrica International Scientific Publisher Monza/Italy

‘Phraseologies’ and Italian-English dictionaries: evidence for a proposal

Stefania Nuccorini – Roma Tre University

1. Introduction

This contribution deals with the lexicography-oriented analysis of a few pairs of English and Italian supposed true friends, i.e. semantically cognate words which are used in different syntagmatic patterns in the two languages: as a result the supposed true friends analysed cannot be considered as translational equivalents.

Usually the phraseological differences concerning these pairs of words are not, or are not clearly signalled in bilingual dictionaries and the consequent pitfalls for, in particular, advanced learners engaged in production tasks can be easily imagined. Sometimes considerable semantic shifts in the use of supposed true friends are not signalled either, as clearly shown by the example reported in Partington (1998: 53) concerning *TO SANCTION* (meaning ‘approve’) and *SANZIONARE* (meaning ‘penalise’). Generally speaking, in order to meet advanced learners’ needs a recent proposal has highlighted how a new type of electronic bilingual learner’s dictionary can be produced by linking

> the kind of information given in existing monolingual learner’s dictionaries to the relevant elements from a standard bilingual lexicographic database. (Bogaards and Hannay 2004: 474)

A “new concept bilingual dictionary”, an L1-L2-L2 (Hebrew-English-English Dictionary) has proved extremely well-received by and useful to students engaged in a translation task (Laufer and Levitzky-Aviad 2006). It is argued, though, that most of the relevant phraseological information concerning supposed true friends is not or is not clearly presented in monolingual learner's dictionaries either: thus it is proposed to include it in bilingual dictionaries.[1]

This paper is divided into four main sections: the research background; the methodology used in the analyses carried out in previous case studies and in this contribution; the main findings emerged so far and related to the present proposal; concluding remarks.

2. Background

In the title of this contribution the word 'phraseology', notably in the plural, is used in a way which, though by now rather common particularly in the area of corpus-based or corpus-driven studies, differs considerably from its long-established, Eastern-tradition-derived sense of 'the set of fixed expressions (idioms, proverbs, etc.) of a language'. This is the sense that has long been and still is traditionally associated with lexicography, as testified by a recent project discussing criteria for entry selection in a bilingual phraseological dictionary whose macrostructure is composed of fixed expressions of various nature (Starko 2006). On the other hand 'phraseology' is here used to refer to the phraseological environment of a word, "the whole range of co-occurrence patterns" of a given lexeme (Granger 2005: 167). The word 'pattern' itself refers to the repeated syntagmatic co-occurrence of a lexeme with other lexical or grammatical words: a pattern can be identified by examining the collocational and colligational profile of a word, its semantic preferences and its semantic prosody (Sinclair 1996, 2002 among others; Tognini-Bonelli 2001). In other words 'phraseology' refers to and summarises the "prototypical syntagmatic patterns with which words in use are associated" (Hanks, 2004: 87).

Patterns do not necessarily form collocations: users unaware of them and of their communicative role are likely to produce inappropriate, infelicitous or wrong combinations. In many cases

[1] In a paper presented at the International Informal Colloquium on Possible Dictionaries (Rome, July 6-7, 2007) held at the Department of Linguistics of Roma Tre University, the author of the present article has proposed to compile a dictionary of 'phraseologies' which, in analogy with dictionaries of collocations, should report the contrastive phraseologies of supposed true friends in the microstructure.

the inappropriateness, infelicity or wrongness is due to contrastive elements in the phraseologies of a given lexeme and its equivalent in the foreign language, English in this case. This is particularly true, and extremely relevant both linguistically and lexicographically, in the case of supposed true friends: often their respective phraseologies do not match their lexico-semantic friendship (Partington 1998; Tognini-Bonelli 2001; Nuccorini 2006).

Language learners are usually made aware by teachers of the existence of false friends, and reminded about their relevance to production tasks in didactic materials and in dictionaries, often in special units or sections. Semantically and etymologically cognate words that do not constitute pairs of false friends do not normally receive any attention: by default, in a sense, they are considered as true friends and therefore unproblematic. The pairs analysed, though, have proved highly problematic largely because of contrastive phraseological elements, hardly ever accounted for in dictionaries. Given their contrastive status, which limits their relevance to the language-specific characteristics of Italian and English, the problematic aspects of the pairs considered are usually not included, and rightly so, in monolingual learner’s dictionaries. Thus the new type of dictionary proposed by Bogaards and Hannay, though extremely useful in other respects, most probably would not solve the specific problems illustrated here. Information about the different phraseological use of supposed true friends should be reported in bilingual dictionaries on account not only of their translational purposes but also of the learners’ favour they enjoy since they are very often used in activities other than translation tasks.

It is well known from various surveys into dictionary use and from descriptive and analytical comparisons between monolingual and bilingual dictionaries (among others, Atkins 1985; Atkins and Knowles 1990; Nuccorini 1992, 1994; Atkins and Varantola 1997; Rundell 1999; Béjoint 2002) that bilingual dictionaries tend to be overused and misused: they are used more than, sometimes instead of, monolingual, especially learner’s, dictionaries despite teachers’ and material writers’ recommendations. The different purposes and as a consequence the different performance of the two typologies of dictionaries are strictly connected with the different advantages and disadvantages in the look-up processes. Though on the one hand

> dictionary skills must be taught carefully and thoroughly, if dictionary users are to extract from their dictionaries the information which lexicographers have put into them (Atkins and Varantola 1997: 36)

the matter-of-fact use of dictionaries, which applies to bilingual dictionaries in particular, has yielded the following conclusion:

> a more realistic strategy is to aim for dictionaries whose structure is so transparent that students do not need to *learn* how to use them, and whose content is presented in such straightforward terms that students will have no difficulty in grasping it. (Rundell 1999: 48)

It is predominantly in this perspective that ‘better’ bilingual dictionaries should be produced: they should aim at meeting users’ real needs, as clearly exemplified by Béjoint (2002) with his proposal for a bilingual dictionary for comprehension. In writing tasks learners tend to mentally translate their thoughts from Italian into English and they turn to bilingual dictionaries for help. Bilingual learner’s dictionaries were “conspicuously absent” in 1987 (Snell-Hornby 1987: 159) and bilingual lexicography was still deemed “strikingly immobile” in 2004 (Bogaards and Hannay 2004: 463). Indeed bilingual lexicography is constantly looking ahead: ‘towards’ and ‘dream’ are words still very frequently used in many academic papers (among others, Iamartino 2006,[2] Morini 2006). Notable exceptions to this picture are the Van Dale Dictionaries (“miracle lexicographique”, Hausmann 2002: 11) and a few proposals and achievements among which the already cited Boogards and Hannay model (2004) for a bilingual + monolingual dictionary and the Laufer, Levitzky-Aviad L1-L2-L2 BD+ (2006): the critical perspective of the former and the user’s perspective of the latter offer great insights into the area which Hartmann (forthcoming) has recently referred to as “reference science”. The Bilexicon Project (Siepmann 2006) and electronic dictionaries (Lo Cascio 2005 for English and Dutch) seem to offer more and more refined tools for both learners and translators.

Bilingualized dictionaries would not be an appropriate answer to the problems of equivalence and use analysed in this paper because

[2] In his paper Iamartino refers to Johnson’s ‘dream’ to describe all the words of the English language (Iamartino 2006: 101).

they are based on learner’s monolinguals and because, as Marello points out (1998: 310), they are “essentially readers’ dictionaries, giving various degrees of support for the decoder”. The proposal for the inclusion of information about the phraseologies of supposed true friends in bilinguals, in addition to the information reported in monolingual learner’s dictionaries, is meant for production tasks.

3. Methodology

A few supposed true friends of Latin origin and an occasional example of loan words from a third language will be analysed both from the linguistic and from the lexicographic point of view. Morphological look-alikes and calques will be excluded for different reasons. In the first case because pairs of words such as *TASK* and *TASCA* are not semantically related; in the second because Italian calques of English words often concern new uses of already existing cognate words in Italian: the ‘new’ uses turn previous false friends into true friends, as in the case of *REALIZZARE* which has borrowed from *REALIZE* the sense ‘to be aware of something’s importance’. On the other hand the pairs of words analysed in this paper concern true friends which have turned out to be false friends.

The pairs of words discussed here have already been analysed in previous case studies, following the methodology expounded in Tognini-Bonelli (2001), with particular reference to the search for functional equivalents, i.e. lexemes showing equivalence of function in their contexts of use. Data from comparable corpora (see note 3) has shown how the supposed true friends analysed here are not functionally equivalent as their phraseologies often diverge. In the present paper, their (mis)use has been checked against the Italian component of the International Corpus of Learner English (ICLE-IT) (Granger, Dagneux, Meunier 2002; Prat Zagrebelsky 2004) for expected phraseological errors. The lexicographic treatment of the pairs discussed has then been analysed in two bilingual dictionaries (Picchi 1999, Oxford-Paravia 2001): both dictionaries present themselves as ‘bidirectional’ in Marello’s terminology (1989). The treatment of the English members of the pairs has also been analysed in two monolingual learner’s dictionaries (COBUILD 2003; OALD 2005) to see whether and to what extent the information offered could be sufficiently and efficiently added to that reported in the bilinguals.

The limits of the data and of the findings on which this paper is based must be accounted for: the corpora used for the analysis of the pairs discussed concern the written language only;[3] many other dictionaries could have been consulted; results concern case studies. However, as will be illustrated, findings show that there is sufficient and sufficiently relevant evidence for the inclusion of information about contrastive phraseologies in bilingual dictionaries. The pairs discussed, which represent different types of contrastive phraseologies, are the following:

- *ASSOLUTAMENTE/ABSOLUTELY*;
- *REALE/REAL*;
- *TERRORISTA/TERRORIST*;
- *KAMIKAZE/KAMIKAZE*.

4. Findings

ABSOLUTELY/ASSOLUTAMENTE

The phraseology of the pair of adverbs *ASSOLUTAMENTE/ABSOLUTELY* has been analysed in particular in Partington (1998) and in Nuccorini (2006). Most relevant differences, and therefore probable causes of errors, concern their use in (morphologically) negative co-texts.[4]

[3] Data concerning the following pairs, *ASSOLUTAMENTE/ABSOLUTELY*, *TERRORISTA/TERRORIST* and *KAMIKAZE/KAMIKAZE*, have been drawn from the sub-corpus, called 'the press', of the Corpus di Italiano Scritto (CORIS) and from the qualitatively and quantitatively analogous British English newspaper and magazine subcorpora of the on-line version of the Bank of English (see Nuccorini 2006, for details). Data concerning REALE/REAL come from the Birmingham Corpus of written English and from the Corpus of Contemporary Italian (see Tognini-Bonelli 2001, for details). In addition, Tognini-Bonelli also used data from the BBC Corpus within the Bank of English, a corpus of spoken (written-to-be-spoken) language.

[4] There are differences in the use of the two adverbs in positive co-texts as well, but these will not be analysed in this paper.

ABSOLUTELY is very frequent in the 'no-negation' system,[5] a pattern which has no real equivalent in Italian. This would probably lead to the under-use of this typical construction. Conversely, as a sentence adverb *ABSOLUTELY* is not used in the 'not-negation' system, whereas the use of *ASSOLUTAMENTE* in this pattern is very frequent in Italian: for example a sentence such as "[…] non corrisponde assolutamente alle aspettative", would be wrongly translated into *"[it] does not match absolutely expectations". As an adjective modifier *ABSOLUTELY*, unlike *ASSOLUTAMENTE*, is rarely used in negative predicative constructions: in these cases it is used just with a restricted group of adjectives including *sure, certain*, *clear*, *necessary, true* and a few others only.

In ICLE-It errors concerning the misuse of *ABSOLUTELY* concern the latter case in particular: examples include *"it isn't absolutely up to date"; *"if I'm absolutely not interested in a subject", *"this is not absolutely right".

In Picchi (1999) *ABSOLUTELY* is the first equivalent reported in the first meaning[6] in the entry for *ASSOLUTAMENTE*. It is used in one example only (the third) in a positive co-text. Three examples rightly report the typical use of *ASSOLUTAMENTE* as a sentence adverb in the Italian not-negation system. One of these concerns the first meaning: the adverb used is "totally" ("non era assolutamente d'accordo con lei"; "he totally disagreed with her"), but no explicit information rules out the use of *ABSOLUTELY* in this case (see note 7). In the other two examples *ASSOLUTAMENTE* is rendered as "possibly", the only equivalent recorded for the second meaning: "non posso assolutamente dirtelo" is translated into "I can't possibly tell you" and "non possono assolutamente permettersi un'auto" into "they can't possibly afford a car".

ASSOLUTAMENTE is the first equivalent recorded in the entry for *ABSOLUTELY*. It is never used in the examples, not even in those in which it is a functional equivalent of its cognate English adverb:

[5] In English there are two competing systems of negation, the not-negation system and the no-negation system, exemplified by 'not to know anything' as opposed to 'to know nothing'.

[6] *Stricto sensu* bilingual dictionaries do not report meanings. The term 'meaning' is here used in a general sense to refer to the numbered sections, often based on meaning discriminators, in the entries.

"I absolutely agree" is translated into "Sono totalmente d'accordo". The phraseological differences between the English and the Italian cognate adverbs are best symbolised in this and in the previous example, which capture the complex contrastive relation between them in negative co-texts. Thus while "I absolutely agree" is a perfect functional equivalent of "sono assolutamente d'accordo", the same does not hold true for the corresponding negative statements. "Non sono assolutamente d'accordo" cannot be translated into *"I don't absolutely agree" nor into *"I absolutely disagree".[7]

No example illustrates the typical use of *ABSOLUTELY* in the no-negation system.

In the Oxford-Paravia dictionary too *ABSOLUTELY* is the first equivalent of *ASSOLUTAMENTE* in the first meaning ("necessariamente, ad ogni costo"), and in the second one ("del tutto"), but it is never used in the translations of the examples. It is also the second equivalent reported in the third meaning ("per niente"). Notably *ABSOLUTELY* is not used either in the translations of the examples in which it is a functional equivalent of its cognate Italian adverb: as an adjective modifier in negative predicative constructions with the adjective *SURE* with which it is typically used ("non sono assolutamente sicuro" is translated into "I'm not awfully sure"); as a sentence adverb ("non hai fatto assolutamente nulla per fermarli" is translated into "you did nothing at all to stop them"). In the latter case its typical use in the no-negation system could have been highlighted. Three examples illustrate the typical use of *ASSOLUTAMENTE* as a sentence adverb in the Italian not-negation system: it is never rendered as *ABSOLUTELY*, but here too, no other explicit information is recorded.

In the entry for *ABSOLUTELY, ASSOLUTAMENTE* is the first equivalent recorded though it is never used in the examples, none of which concerns the typical use of the English adverb in the no-negation system.

OALD emphasises the use of *ABSOLUTELY* in the no-negation system in the second definition. Obviously it also reports other relevant information concerning the use of the English adverb in

[7] There is no occurrence of *ABSOLUTELY* with *DISAGREE* in the Bank of English. The English adverb occurs with verbs expressing denial (*REFUSE*, rarely *DENY*) but not with morphologically negative ones, such as *DISAGREE*.

positive co-texts which would be extremely useful in a contrastive perspective.

COBUILD reports one example of *ABSOLUTELY* in the no-negation system. Other information concerns its use in positive co-texts both as a sentence adverb and as an adjective modifier.[8]

REALE/REAL

The pair *REALE*/*REAL* has been analysed in Tognini-Bonelli (2001): the author first examines *REAL* and then *REALE* considered as a *prima facie* equivalent of the English adjective. In the sense ‘existing in reality’ they share the same typical abstract nouns as general collocates and a few specific ones from the financial area. In the latter case they show considerable symmetry, whereas in the former phraseological, colligational differences are more relevant. For example the modifier *VERY* is rather frequent in English (“a very real need”) whereas usually *REALE* is not modified.

The ‘real problems’ concern the delexical functions of *REAL,* when the adjective is not used in the sense ‘existing in reality’. The delexicalised focusing and selective functions of *REAL*, examined by Tognini-Bonelli in depth, can be summarised as follows. The focusing function of *REAL*, typically preceded by the indefinite article, emphasizes “certain characteristics already present in the noun” (Tognini-Bonelli 2001: 118): “a real problem” is a serious problem. In Italian this function is not realised by *REALE* but by *VERO* usually in the marked attributive position: “un vero problema”. Most interestingly, according to Tognini-Bonelli’s data, the focusing function of *REAL* accounts for 80% of its occurrences, while *VERO* is used in its matching focusing function in only 11% of its occurrences: this leaves ample space for further research.

In its selective function the English adjective, usually significatively preceded by the definite article, conveys some type of contrast: for example in “the real problems” it implies reference to other possible (less serious or less important) problems. Tognini-Bonelli (2001: 147) argues that *REALE* has no such function: to express it, Italian uses once more the adjective *VERO*, usually, but

[8] For a more detailed analysis of the treatment of *ABSOLUTELY* in learner’s dictionaries see Nuccorini 2007.

not obligatorily, in the pre-nominal, marked position: “i veri problemi”. In these cases though, the use of *REALE*, in its usual post-nominal position, is also possible: “i problemi reali”. However the unmarked position of *REALE* also allows its literal, lexical interpretation, i.e. “the problems existing in reality”. Further research is necessary to shed light on this aspect and on the definition of the web of functionally equivalent English and Italian words triggered off by the analysis of *REAL* and *REALE*.

In ICLE-IT *REAL* is a rather frequent word. Most errors concern its use with wrong or unusual collocates: *“real cover girl”, *“real programmed instrument”, *“real ideal”, *“real culture”, ?“real initiatives”, ?“real quarrels”. Strangely in most of these erroneous combinations *REALE* could not be used either. *REAL* is hardly ever used in its focusing function with the indefinite article, as expected, whereas it is often used in its selective function with the definite article.

In the entry for *REALE*[9] in the Picchi dictionary *REAL* is listed among other equivalents in the first meaning (“vero”) and in the third (“tangibile, concreto”). It is the only equivalent reported in the second meaning concerning the specialist use of the adjective in the legal, financial, economic and mathematical sectors. The English adjective is used in almost all the translated examples, notably the first one: “problemi reali”, “real problems”.[10] Among those listed, the only other equivalents used in a very few examples are “true” and “concrete”, with no specific indication.

The entry for *REAL* is considerably longer than that for *REALE*, thus signalling even more specific uses of the English adjective not covered by its cognate Italian equivalent. *REALE* is listed together with other adjectives in the first meaning (“genuine”), which refers mostly to the focusing function of *REAL*. It is used in one surprising example only, “real number”, “numero reale”. It is also recorded in the third and fourth meanings, mostly concerning economic and financial expressions, and almost always used in the examples. It is

[9] The entries for the adjective meaning ‘royal’ and for the noun are not relevant to the present analysis.

[10] The lack of the (definite) article in this example and its reference to the meaning ‘vero’ reinforce the lexical interpretation of the Italian adjective, as opposed to the delexicalised, selective function of the English one.

not given among the equivalents for the second meaning (“proper, true”) but interestingly it is used in the translation of the example “real problems” together with “effettivi”: this too seems to point to the lexical use of *REALE* in this case.

In the Oxford-Paravia dictionary *REAL* is the only equivalent reported for four meanings (mostly concerning specialist uses) of *REALE*. The first meaning is divided into two sub-senses: “non immaginario”, for which *REAL* is given as the only equivalent, and “concreto” for which the dictionary also reports “true” and “actual”.

In the entry for *REAL*, *REALE* is recorded only in the first meaning “actual, not imaginary or theoretical”. One of the examples is “the threat is very real” translated into “la minaccia è molto reale”: the use of “molto” sounds unusual (other adverbs, such as “decisamente” or “veramente” could be used instead). In all the other meanings, among which the fourth one concerns the emphatic (or focusing) function of the adjective, the dictionary reports other Italian equivalents. Several combinations (for example “real number” or “real time”) are considered as compounds and listed as headwords.

OALD reports four general ‘short-cuts’ or meanings for *REAL*, each followed by definitions and examples: “existing/not imagined”; “true/genuine”; “for emphasis”; “money/income”. Each refers to, respectively, the lexical meaning of the adjective; its selective function; its focusing function; and its specialist uses. This picture confirms, on the basis of data taken from a very large corpus, Tognini-Bonelli’s findings about the typical phraseological profile of *REAL*.

COBUILD reports ten definitions which explain subtle differences in the lexical and delexical uses of the adjective in detail. Definitions 7, 8 and 9 in particular concern the focusing function of *REAL*, definition 10 its specialist uses. Some expressions are listed as compound headwords: “real life”, “real time”, “real world”.

TERRORISTA/TERRORIST

This pair of cognate nouns has been analysed in Nuccorini (2006). Relevant differences concern their respective frequencies in use and their association with socio-historical and culture-specific 'national' phenomena. *TERRORISTA* is much more frequent than *TERRORIST* mostly used in its adjectival, pre-nominal function. Unlike the English noun, the Italian one is often used as a general term, as are the plural forms of both. In the singular they are both modified by adjectives of nationality or locatives or relatives pointing to international situations. However, when concerning English or, conversely, Italian situations, the two nouns are typically associated with different scenarios: the IRA on the one hand and, on the other, left or right extremists or the mafia. Thus, in these cases, the two nouns are not functional equivalents: English *BOMBER* is frequently used[11] in most cases in which Italian uses *TERRORIST* or *KAMIKAZE*, as will be seen.

There are no occurrences of *TERRORIST* or *TERRORISTS* in ICLE-It, most probably because the noun is not associated with the topics discussed in the students' essays included in the corpus.

In the Picchi dictionary *TERRORIST* is the only equivalent recorded in the entry for *TERRORISTA*, labelled as noun and as adjective. Very interestingly, in the entry for *TERRORIST* the only example is "IRA terrorists".

In the Oxford-Paravia dictionary *TERRORISTA* and *TERRORIST* are the only equivalents of each other reported in their respective entries. No example is reported. The English headword is labelled as noun and modifier.

OALD defines *TERRORIST*, labelled as noun, as "a person who takes part in terrorism" and reports one example only showing the 'general' use of the noun in the plural.

COBUILD includes references to 'murder' and 'bombing' in the definition of *TERRORIST*, labelled as noun, and reports one example concerning its adjectival use.

[11] With specific reference to the July 2005 London bombings, the Gilbert & George exhibition at The New Tate in February 2007 showed what looked like a collection of a considerable quantity of headlines always including '(suicide) bomber(s)'.

KAMIKAZE/KAMIKAZE

The Japanese word has been borrowed both in English and in Italian: on this basis the Italian and the English loan words are considered as *prima facie* equivalents of each other in Nuccorini (2006).

In Italian KAMIKAZE is used predominantly as a noun, in particular in its extended sense of ‘suicide bomber’ (i.e. a ‘terrorist’), a sense not present in the original Japanese word. It is also used figuratively as a modifier in the fields of sport and politics in particular. In English *KAMIKAZE* (the pronunciation varies in the two languages) is an adjective only, used in its original Japanese sense to modify nouns such as ‘pilot’ or ‘mission’. It is occasionally used figuratively in the field of sport. The use of the two words as different parts of speech and their collocational profiles, in a word their diverging phraseologies, have turned their supposed true friendship into false friendship. The analysis of other words posited as possible equivalents has shown that *SUICIDE BOMBER* is the English functional equivalent of Italian *KAMIKAZE* (and vice-versa).

As in the case of *TERRORIST* there is no occurrence of *KAMIKAZE* in ICLE-IT.

In the Picchi dictionary Italian *KAMIKAZE* is labelled as adjective and as noun: the sole equivalent given is “kamikaze”.

In the English-Italian part *KAMIKAZE* is labelled as a countable noun and as an adjective: the equivalent given in the first case is “kamikaze”, while in the second the equivalents are the adjectival phrase “da kamikaze” and the adjective “suicida” used in the translated example. The last example is revealing: “a kamikaze pilot” is translated into “un (pilota) kamikaze”.

In the Oxford-Paravia dictionary the Italian word is labelled as adjective and as noun: the only equivalent is “kamikaze”.

The treatment of the English headword, labelled as noun and adjective, is analogous to that in the other bilingual dictionary, but no example is offered.

OALD labels *KAMIKAZE* as pre-nominal adjective and offers a definition which is clearly related to its original Japanese use, and two examples, one of which about its figurative use. The synonym “suicidal” is reported at the end of the entry.

COBUILD as usual reports information about part of speech (pre-nominal adjective) in the extra-column. The definition of *KAMIKAZE* includes some reference to its use in possible contexts other than the original one ("if someone such as a soldier or a terrorist performs a kamikaze act …"): the only example reports the typical use "kamikaze pilot".

5. Conclusion

The contrastive aspects concerning the profiles of English and Italian supposed true friends and/or *prima facie* equivalents analysed so far constitute case studies offering evidence of different types of phraseological mismatches with different lexicographic relevance.

In the case of the use of *ASSOLUTAMENTE*/*ABSOLUTELY* restrictions and constraints concern predominantly the English adverb. Conversely in the case of *REALE*/*REAL* most use restrictions and constraints concern the Italian adjective. Both *TERRORISTA* and *TERRORIST* show contextually specific uses in some relevant cases. The typically divergent use of English and Italian *KAMIKAZE* allows to consider them as false friends.

For each pair the lexicographic treatment has proven unsatisfactory. The presence of *ABSOLUTELY*, hardly ever used in the translated examples, as the first equivalent of *ASSOLUTAMENTE* is misleading. More information about the typical use of the English adverb, (partially) present in the learner's dictionaries analysed, would be necessary. The differences in function between *REALE* and *REAL*, though sometimes difficult to pinpoint, are not reported and some confusion concerning "vero" as a synonym of "reale" adds to the shortcomings of their treatment which is much clearer in the monolingual dictionaries analysed. No relevant information, but for the Picchi example, is offered in the entries for *TERRRORISTA* and *TERRORIST* and in this case neither OALD nor COBUILD are helpful. The analysis of the lexicographical treatment of English and Italian *KAMIKAZE* speaks for itself.

In the perspective of Rundell's proposal, with particular reference to the area of contrastive phraseologies, much remains to be done. The necessary background linguistic analysis should move from case studies to a systematic approach and include the description of alternative equivalents. As Tognini-Bonelli suggests (2001: 149)

> A data-base of translation equivalents will have to develop a system of annotation capable of accounting for […] formal and functional differences between source and target language.

The examples analysed offer enough evidence for further systematic research and for the inclusion of contrastive phraseologies, particularly those concerning supposed true friends, in bilingual (learner's) dictionaries.

References

Atkins, B.T. (1985) "Monolingual and bilingual learner's dictionaries: a comparison", in R. Ilson (ed.) *Dictionaries, Lexicography and Language Learners*, Pergamon Press, Oxford, pp. 15-24.

Atkins, B.T. and K. Varantola (1997) "Monitoring dictionary use", *International Journal of Lexicography* 10 (1), pp. 1-45.

Atkins, B.T. and F.E. Knowles (1990) "Interim report on the EURALEX/AILA research project into dictionary use", in T. Magay and J. Zigány (eds), *BudaLEX '88 Proceedings*, Akadémiai Kiadó, Budapest, pp. 381-392.

Béjoint, H. (2002) "Towards a bilingual dictionary for 'comprehension'", in E. Ferrario and V. Pulcini (eds) *La lessicografia bilingue tra presente e avvenire*, Edizioni Mercurio, Vercelli, pp. 33-48.

Bogaards, P. and M. Hannay (2004) "Towards a new type of bilingual dictionary" in G. Williams and S. Vessier (eds), *11th EURALEX Congress Proceedings*, UBS, Lorient, pp. 463-474.

Collins COBUILD Advanced Learner's English Dictionary (2003) Harper Collins Publisher, London.

Granger, S., E. Dagneux and F. Meunier (eds) (2002) *International Corpus of Learner English*, UCL Presses Universitaires de Louvain, Louvain.

Granger, S. (2005) "Pushing back the limits of phraseology: how far can we go?", in C. Cosme, C. Gouverneur, F. Meunier and M. Paquot (eds), *Phraseology/ Phraséologie. Abstracts of the Conference Papers, 13-15 October 2005*, Centre for English Corpus Linguistics, Louvain-la-Neuve, pp. 165-168.

Hanks, P. (2004) "Corpus pattern analysis", in G. Williams and S. Vessier (eds), *11th EURALEX Congress Proceedings*, UBS, Lorient, pp. 87-97.

Hartmann, R.R.K. (forthcoming) "Promoting interdisciplinary collaboration between lexicology, lexicography, terminology and translation: towards reference science?", paper presented at the International Conference on *Lexicology and Lexicography of Domain-specific Languages*, Palermo, June 21-23, 2007.

Hausmann, F.J. (2002) "La lexicographie bilingue en Europe: peut-on l'améliorer?", in E. Ferrario and V. Pulcini (eds), *La lessicografia bilingue tra presente e avvenire*, Edizioni Mercurio, Vercelli, pp. 11-31.

Iamartino, G. (2006) "Dal lessicografo al traduttore: un sogno che si realizza?", in F. San Vicente (ed.), *Lessicografia bilingue e traduzione: metodi, strumenti, approcci attuali*, Polimetrica, Monza, pp. 101-132.

Laufer, B. and T. Levitzky-Aviad (2006) "Examining the effectiveness of 'Bilingual Dictionary Plus' – a dictionary for production in a foreign language", *International Journal of Lexicography* 19 (2), pp. 135-155.

Lo Cascio V. (2005) *Grande dizionario elettronico italiano-neerlandese, neerlandese-italiano*, Fondazione Italned, Amstelveen.

Marello, C. (1989) *Dizionari bilingui*, Zanichelli, Bologna.

Marello, C. (1998) "Hornby's bilingualized dictionaries", *International Journal of Lexicography* 11 (4), pp. 293-314.

Morini, M. (2006) "Il dizionario del traduttore. Un sogno che si realizza?", in F. San Vicente (ed.), *Lessicografia bilingue e traduzione: metodi, strumenti. approcci attuali*, Polimetrica, Monza, pp. 165-179.

Nuccorini, S. (1992) "Monitoring dictionary use", in U. Tonnola, K. Varantola, T. Salmi-Tolonen and J. Schopp (eds), *EURALEX '92 Proceedings*, Studia Translatologica, University of Tampere, Tampere, pp. 89-102.

Nuccorini, S. (1994) "On dictionary misuse", in W. Martin, M. Meijs, E. Moerland, E. ten Pas, P. van Sterkenburg and P. Vossen (eds), *EURALEX '94 Proceedings*, Vrije Universitat, Amsterdam, pp. 586-597.

Nuccorini, S. (2006) "In search of phraseologies: discovering divergences in the use of English and Italian true friends", *European Journal of English Studies* 10 (1), pp. 33-47.

Nuccorini, S. (2007) "Note su alcune 'fraseologie' nei dizionari pedagogici inglesi più recenti", in N. Minerva (ed.) *Lessicologia e lessicografia negli insegnamenti linguistici*, Cooperativa Libraria Universitaria Editrice, Bologna, pp. 123-139

Oxford Advanced Learner's Dictionary (2005) Oxford University Press, Oxford.

Oxford-Paravia, Il dizionario inglese-italiano, italiano-inglese (2001), Paravia, Torino.

Partington, A. (1998) *Patterns and Meanings*, John Benjamins, Amsterdam.

Picchi, Fernando (1999) *Grande dizionario inglese-italiano/italiano-inglese*, Hoepli, Milano.

Prat Zagrebelsky, M.T. (2004) (ed.) *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage,* Edizioni dell'Orso, Alessandria.

Rundell, M. (1999) “Dictionary use in production”, *International Journal of Lexicography* 12 (1), pp. 35-53.

Siepmann, D. (2006) “Collocation, colligation and encoding dictionaries. Part II: lexicographic aspects”, *International Journal of Lexicography* 19 (1), pp. 1-39.

Sinclair, J.M. (1996) “The search for units of meaning”, *Textus* IX (1), pp.75-106.

Sinclair, J.M. (2002) “Phraseognomy”, in S. Nuccorini (ed.), *Phrases and Phraseology: Data and Descriptions*, Peter Lang, Bern, pp. 17-26.

Snell-Hornby, M. (1987) “Towards a learner’s bilingual dictionary”, in A.P. Cowie (ed.) *The Dictionary and the Language Learner*, Niemeyer, Tübingen, pp. 159-170.

Starko, V. (2006) “Entry selection for a bilingual phraseological dictionary” in E. Corino, C. Marello and C. Onesti (eds), *Proceedings XII EURALEX International Congress*, Edizioni dell’Orso, Alessandria, pp.1045-1054.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work,* John Benjamins, Amsterdam.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 189-203 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Corpora and lexicography: the case of a dictionary of Anglicisms

Virginia Pulcini – University of Turin

1. Introduction

Computer Corpus Lexicography has greatly contributed to a new perspective in the study of lexis and in the practice of dictionary-making (Ooi 1998; Sinclair 2003). However technically complex some of the literature in this area may be, and in spite of scepticism on the part of some old-school linguists, modern lexicography cannot do without the lexical information that corpora can provide. The good news is that today corpora have become much more accessible and user-friendly than they were a decade ago, so that it is no longer indispensable for lexicographers to be experts in the computing dimension of corpus linguistics to access corpora and extract the lexical information that they need to compile a dictionary.

Ever since the pioneering COBUILD project set up by John Sinclair (Sinclair 1991), the 'corpus revolution' has deeply changed the principles and methods of dictionary-making. Evidence from real language use can provide comprehensive data on the lexical profile of words, including their frequency, register, range of meanings, collocations, phraseology and examples, as well as objective support to lexicographers' intuitions. Most English dictionaries published in the past two decades have claimed to be 'corpus-based', although in many cases the reference corpora may be only accessible to insiders and not to the whole community of scholars. This is motivated by the huge financial investments that publishers make to build their own databases and support their dictionary projects, as well as by copyright restrictions.

Today many corpora have been made available for free online query, or are accessible by a reasonable subscription fee; the query procedures have also become user-friendly, so that the extraction of information is an easy and even enjoyable activity for scholars and students who are not conversant with the technicalities of corpus linguistics.

This paper describes how corpora are being exploited for the compilation of a *Dizionario di anglicismi* which is currently underway. The corpora are the La Repubblica Corpus, the itWaC and the BNC. The project of this dictionary, already described elsewhere in detail (Pulcini 2006), has consisted in the building of a large database of Anglicisms taken from already existing lexicographic sources (Görlach 2001; De Mauro and Mancini 2003; Zingarelli 2006) and the subsequent refinement of the wordlist through *ad hoc* criteria of inclusion and exclusion especially by means of corpus evidence (Pulcini, forthcoming). The corpus work consists in the search for each candidate entry in Italian and English corpora, the evaluation of its frequency and the observation of its lexical behaviour in context. In the following sections, the lexical information derived from frequency data and the KWIC format of Anglicisms on both Italian and English corpora will be illustrated.[1]

2. The Corpora

The La Repubblica Corpus contains 380M words from articles published by the Italian daily newspaper between 1985 and 2000. Recorded on 16 CD-roms issued by the newspaper, this corpus was compiled at the *Scuola Superiore di Lingue Moderne per Interpreti e Traduttori* (SSLMIT) in Forlì for research and teaching purposes.[2] As the authors themselves point out, this corpus is very large but it is representative of one and the same register, i.e. written newspaper language from the same source (Baroni *et al.* 2004). In spite of this pitfall, the corpus is by all means suitable for looking

[1] As described in Pulcini (2006) some preliminary corpus analysis was carried out on the HF corpus (a 19.47 million words of newspaper articles from *la Stampa*, *la Repubblica* and il *Corriere della Sera*. (Furiassi and Hofland 2007) and the NUNC (Newsgroups UseNet Corpora) a 237 million word of newsgroup interactions, freely available and queriable online.

[2] See the web page: http://sslmitdev-online.sslmit.unibo.it/corpora/query

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

up Anglicisms, since newspapers record up-to-date national and world news in a wide range of fields and topics that are relevant to modern life and current events. In short, newspapers mirror (and also 'influence') the interests and the culture of the average educated reader, keeping up with facts and innovations. This feature would fit the profile of the target user of the planned dictionary, i.e.

> an educated Italian speaker who is familiar with the language of the mass media and also possesses a certain proficiency in the English language. (Pulcini 2006: 316).

On the other hand, newspapers are also full of short-lived neologisms or *hapax legomena*, of Anglo-American origin in particular, because of the attractiveness of English words and the influence of the English-speaking news media on the Italian press. In our experience, this has proved particularly true for the field of business and finance.

The itWaC (Italian Web as Corpus) is a very large corpus (about 2 billion words), containing texts from Italian web pages. It is part of a suite of corpora which can be accessed through the Sketch Engine program.[3] Besides its size, the advantages of a web-based corpus are that it contains truly up-to-date vocabulary, as well as topics, genres and registers which may not be found in other corpora, such as highly technical terminology or personal, communicative styles, e.g. blogs (Baroni and Ueyama 2006). Although English is no longer the only dominant language of the World Wide Web (Crystal 2003), Anglo-American culture and technology are still certainly so. Therefore English dominates in computer terminology and music, films, books, videogames and in a wide range of products described and advertised on the web. Moreover, web pages normally contain several links to other web pages which may be in English, a phenomenon which should be avoided in the building of a corpus (a process known as 'web crawling') but in fact it is difficult to eliminate. Other pitfalls of web corpora which have been encountered in the course of our corpus query, is that the same web pages may appear many times, so that a search item may obtain, for example, 40 hits which are in fact the same one repeated 40 times. In spite of these problems, a web

[3] The Sketch Engine is a product of Lexical Computing Ltd., directed by Adam Kilgarriff. Access to the Sketch Engine program is granted to registered users through the payment of a license fee.

corpus is indeed an enormous storehouse of lexical information of contemporary, up-to-date, spontaneous, living language (Baroni and Ueyama 2006).

Through the Sketch Engine it is also possible to access the British National Corpus, the 100M word corpus of British English, containing samples of written and spoken language from a wide range of late 20th century sources. The BNC is used to compare the lexical features, meanings and uses of the selected Anglicisms to the lexical features, meanings and uses of the original English word. It was also necessary to establish whether apparently 'English' words are in fact false Anglicisms.

3. Currency and representativeness

Each candidate entry for the *Dizionario di anglicismi* is searched in both Italian corpora and the number of hits is noted down. No frequency scores are calculated, nor is systematic evaluation of differences between the La Repubblica Corpus and itWaC carried out, because the data are only considered as a general index of currency and representativeness. By 'current' we mean "commonly known and accepted in Italian", and by 'representative' we mean that "the word or expression occurs with a certain frequency" (Pulcini 2006; Pulcini, forthcoming). In fact, despite the initial decision to establish a threshold frequency score below which the Anglicism would not qualify as an entry for the dictionary, such a drastic criterion was abandoned. Some specific cases will explain the reason.

The item *AFTERSHAVE* for example, appears in the La Repubblica Corpus (henceforth Rep) only 5 times (0 as *AFTERSHAVE*, 1 as *AFTER-SHAVE* and 4 as *AFTER SHAVE*), and 20 times in the itWaC (2 as *AFTERSHAVE*, 5 as *AFTER-SHAVE* and 13 as *AFTER SHAVE*). In spite of such unexpectedly low presence of this word in both corpora, counterbalanced by 74 occurrences of *DOPOBARBA* in the Rep and 300 in the itWaC, it was decided, on a purely subjective basis as a native speaker of Italian, that the word cannot be excluded from the dictionary because of its currency. As the term is normally written on aftershave bottles and can be seen in advertisements, in fashion magazines and in perfume shops, it is safe to say that most Italians would know it.

An equally subjective decision was taken for ALASKAN MALAMUTE

(4 in the Rep and 12 in the itWaC) and BASS REFLEX (0 in the Rep and 28 in the itWaC, considering different spellings), because the former was judged to be too rare, if not unknown, and the latter too technical.

In sum, in my view, the corpus data are an important index of currency but figures must be balanced with the criteria set up for the dictionary (Pulcini, forthcoming)[4] and possibly with additional information from other sources, e.g. the archives of daily newspapers and magazines (*Il Corriere della Sera*, *l'Espresso*, *Le Scienze*), and in some cases the opinions of Italian experts in certain fields. Last and most importantly, the native lexicographer's judgement will weigh up the different bits of information and make a final decision.

Apart from problematic cases, the two corpora have confirmed expectations, giving equally numerous hits for well-established Anglicisms (e.g. *BABY SITTER*: 829 in the Rep and 3,379 in the itWaC; *AIRBAG*: 1,035 in the Rep and 2,180 in the itWaC)[5] and no hits for rare ones (*ACCROBRANCHING*, *AERO-DANCE*, *BAKING-SODA BEACH MOVIE*, *BLIND BUYING*).[6]

As stated above, no systematic comparison has been carried out between the two corpora, but the much higher number of computer terms in the itWaC is evident. Considering that the itWaC is 5 times larger than the Rep, the following examples show a marked difference between the two corpora in computer terms: *ATTACHMENT* (9 hits in the Rep, 1,044 in the itWaC), *BACKUP* (14 Rep, 8,144 itWaC), *BACKSLASH* (0 Rep, 314 itWaC), *BACKSPACE* (0 Rep, 275 itWaC), *BATCH* (1 Rep, 1,067 itWaC), *BANNER* (23 Rep, 8,104 itWaC), *BIT* (344 Rep, 27,000 itWaC).

While very few cases have been found so far of Anglicisms which are present in the Rep but not in the itWaC, e.g. *BABY*

[4] a) the entries should be part of modern vocabulary (not archaic or obsolescent); b) the entries should be 'recognizable' by an educated Italian user; c) the presence of the entries in Italian should be frequent enough to guarantee a minimum degree of 'assimilation' ('acclimatization'). (Pulcini, forthcoming)

[5] All spellings are considered in the quoted examples (one word, two words separated by a hyphen, two separate words). See below in section 4 for a discussion on orthographic differences.

[6] At this time of writing only the entries starting with 'A' and 'B' of the planned dictionary have been fully completed, therefore the examples quoted will refer to these entries.

PUSHER (6 Rep, 0 itWaC), the opposite case is quite frequent. Besides the numerous computer terms found in the itWaC, and many rare or technical terms which were eventually excluded from the dictionary, we can quote *AQUAGYM* (0 Rep, 90 itWaC)[7] and *BEACH CLUB* (0 Rep, 34 itWaC) among the few inclusions. Because the Rep covers a time span of 16 years up to 2000, no Anglicisms adopted after that date are found, e.g. *BLACK BLOC* (0 Rep, 1,766 itWaC),[8] the name of anarchist protest groups which became active in Italy from the G8 meeting in Genoa in 2001.

To conclude, the data obtained from both corpora are extremely useful to confirm, first of all, that a candidate Anglicism 'is' current in the chosen reference sources. Secondly, for the lexicographer, a high number of hits in both corpora may clear all doubts as to the inclusion of the word in the dictionary, while a low number of hits or its absence calls for further investigation.

4. Compounds and hyphenation

Dictionaries are primarily used for looking up the meaning of a word, but the second most common reason is to check the spelling of a word (Jackson 2002). Orthography is particularly idiosyncratic in English compounds, which may appear as an unbroken orthographic word, a hyphenated word or as two separate words. As Bauer (1988) explains, some compounds are the result of a morphological process, whereby two words are joined together, e.g. *BASEBALL*, to form a new one, usually taking one single stress in initial position; instead, some compounds derive from a syntactic process, whereby a modifying element is joined to a head, e.g. *BOAT PEOPLE*, in which case the elements remain separated and are individually accented. The intermediate case is when the two elements are linked by a hyphen. No safe rule exists and practice varies, with the result of coexisting forms for the same compound, e.g. *BESTSELLER*, *BEST-SELLER*, *BEST SELLER*.[9] As stated by Bauer (1988: 101):

[7] The hybrid *ACQUAGYM* produced 2 hits in the Rep and 287 in the ItWaC.

[8] All the orthographic forms *BLACK BLOC*, *BLACK-BLOCK*, *BLACK-BLOC*, *BLACK-BLOCK*, *BLACKBLOCK* and *BLACKBLOC* are considered.

[9] The BNC produced 111 hits for *BESTSELLER*, 100 for *BEST-SELLER* and 88 for *BEST SELLER*.

> ... it is worth making the point that hyphenation in English is totally random, and does not necessarily prove anything at all about the linguistic status of strings of elements.

Which orthographic form, or how many different ones of the same headword, should be included in a dictionary is another major stumbling block for the lexicographer (Furiassi 2005). To cast light on this unruly aspect of the English language, all candidate compound Anglicisms are looked up in the corpora in the three possible forms: one solid word, a hyphenated word, two separate elements. The dictionary will register all representative spellings found in the corpora, which appear to be fairly consistent in Italian, as is shown in Table 1 below. A comparison with the original English spelling according to the BNC is also included.

	Rep	**itWaC**	**BNC**
all-inclusive	0	107	**34**
all inclusive	**2**	**454**	14
allinclusive	0	5	1
antitrust	**4505**	**5986**	**79**
anti-trust	561	297	56
anti trust	34	40	1
back office	**23**	**732**	**30**
back-office	6	400	11
backoffice	1	107	0
body building	**172**	**637**	**17**
bodybuilding	6	270	17
body-building	37	141	14
benchmark	**57**	**2080**	**285**
bench-mark	0	0	10
bench mark	0	1	18
by-pass	**331**	**694**	299
bypass	167	631	**955**
by pass	121	175	106

Table 1. Orthographic forms of some Anglicisms in the La Repubblica Corpus, the itWaC and the BNC. The most frequent forms are in bold type.

The Italian data appear to be quite consistent: *ALL INCLUSIVE* is more commonly spelt as two separate words, but is also quite frequently hyphenated; *ANTITRUST* and *BENCHMARK* are spelt as solid words but *ANTITRUST* is also found as hyphenated, unlike *BENCHMARK*; for *BACK OFFICE* and *BODY BUILDING*, two separate elements are more common but they are also frequently hyphenated or written as separate elements; *BY-PASS* is more commonly hyphenated but also found as a solid word or separate elements.

If we compare the Italian spellings to the English ones, we notice that there is agreement for *ANTITRUST*, *BACK OFFICE*, *BODY BUILDING* and *BENCHMARK*, but not for *ALL-INCLUSIVE* (the hyphenated form is more common) and *BYPASS* (the solid form is more common than the hyphenated one).

5. Meaning

The Italian corpora are an essential source of information to observe the use of Anglicisms in context in order to detect the grammatical and semantic changes that they undergo as they filter from English into Italian. The semantic aspect of borrowing is a complex phenomenon, "because it involves referential, connotative, contextual, and sociocultural components of meaning." (Pulcini 2002: 162) As a consequence, it is important to check whether the original meanings have been kept, altered, restricted or expanded. Because the *Dizionario di anglicismi* starts from the meanings and senses of Anglicisms in Italian, the next step is to check whether these meanings and senses match those of the original English words, using the BNC. Additional meanings of English words which are not transferred into Italian are not taken into consideration.

Normally, if an English word is borrowed in order to fill a semantic gap in Italian, the referential meaning remains the same. We may quote as examples the tennis term *ACE*,[10] the economic activity called *AGRIBUSINESS*, and the type of accommodation referred to as *BED AND BREAKFAST*. It may rightly be argued, however, that the cultural overtones primed by the mental image of

[10] Apart from the tennis meaning, *ACE* has two additional meanings in English which correspond to the Italian *asso*, i.e. in cards ("the ace of heats", *l'asso di cuori*; an expert, a champion ("the Formula 1 Ace", *l'Asso della Formula 1*).

a typically British ‘bed and breakfast’ may be very different from an Italian one, but from a practical point of view the system is basically the same.

In the case of ‘luxury loans’, the Anglicism is introduced in competition with an already existing word, as in the case of *BLIND DATE* (used less frequently than *appuntamento al buio*) and *BODY BUILDING* (used more frequently than *culturismo*). In these cases, the referential meanings are the same, but the Anglicism is marked by a connotation of style and modernity. Eventually one of the competing words may become obsolete and be dropped (e.g. the old-fashioned *bambinaia* has almost completely given in to the modern *BABY SITTER,* and *pallacorda*, introduced during the Fascist regime to replace *TENNIS*, is by now dead and buried.

Changes in meaning may occur in the form of ‘restriction’, so that the Anglicism carries only one of the several meanings which the original English word possesses (see footnote 10). It is the case of *BENCHMARK*, a financial term in Italian which the forthcoming *Dizionario di anglicismi* defines as:

> indice di riferimento nel mercato finanziario che consente ai gestori e agli investitori di valutare l’andamento positivo o negativo di un fondo di investimento.

While in Italian this is the only meaning associated to the word, in English *BENCHMARK* can also be used generically to refer to “an amount, level, standard etc that you can use for judging how good or bad other things are” (*Macmillan English Dictionary for Advanced Learners*, 2007). So *BENCHMARK* in English may be used generically in different contexts, such as rugby (1), education (2), and science (3), as shown by the following examples taken from the BNC:

(1) Wales' two wins out of eight may be the relevant **benchmark**.

(2) These **benchmarks** would indicate to teachers the things which all children ought to know at a particular stage of their development.

(3) As more breeders being to use cow genetic index as a **benchmark** of genetic potential, they are looking for animals with pedigrees to make significant advances in their herds.

Semantic change may also involve the addition of new meanings or senses with respect to the original English word. An example of this phenomenon is *BACKSTAGE*, which in Italian has two related meanings, defined as follows in the forthcoming *Dizionario di anglicismi*:

> **1a** Documentario dell'attività di preparazione di un film, di un evento mondano, di uno spettacolo teatrale, che ne illustra i problemi tecnici, l'atmosfera, le emozioni e i pettegolezzi. **1b** Area del palcoscenico che si trova dietro le quinte e non visibile dal pubblico, dove gli artisti si preparano prima di entrare in scena e si allestiscono i preparativi per lo spettacolo.

These two meanings are illustrated in (4) and (5) from the itWaC:

(4) Sul secondo DVD, infatti, sono stati inseriti un lungo **backstage**, relativo alla realizzazione di quattro scene chiave del film; un'intervista al regista Giuseppe Tornatore; un'altra allo stesso Tornatore e ad Ennio Morricone riguardo le musiche del film ed una prova d'orchestra, che vi mostrerà in che modo viene incisa la colonna sonora per un film.

(5) Insomma il carrozzone si sta muovendo veloce, il palco sta per prendere vita e qui nel **backstage** si comincia ad avere il mal di pancia, tipico di ogni momento decisivo e importante.

For both meanings, the Anglicism *BACKSTAGE* is used in Italian as a noun preceded by an article, while in English the grammatical functions of *BACKSTAGE* are those of adverb and modifier, as shown in examples (6) and (7) from the BNC:

(6) Later, over a cup of coffee **backstage**, Kylie talks frankly about homesickness and how it had taken her two years to adjust to life in England.

(7) It's the opening night of `Snow White' at the Cambridge Arts Theatre and **backstage** excitement is mounting.

Another instance of the same phenomenon is the word *AFTER HOURS*, which is normally used as a modifier and an adverb in English, while in Italian *AFTER HOURS* has a wider range of uses and senses, as is shown by the following examples: adjective (8), adverb (9), but also noun (10), (11) and (12) meaning 'after-hours party, event or entertainment', 'after hours club' and 'after-hours trading session' respectively. This is caused by the Italian practice of dropping the right-hand element of compounds, creating a false Anglicism which is hardly recognizable by an English native speaker.

(8) Siamo nell'era del trading online, dei call center, dei mercati **after hours**, che aprono quando quelli ufficiali chiudono… (Rep)

(9) Si chiamano smart drinks o power drinks, sono bevande ad alto potere energetico dal contenuto ancora oscuro, bibite reperibili solo nelle discoteche, ormai i coktail preferiti da chi vive **'after hours'**, dall'alba a mezzogiorno. (Rep)

(10) Le '**after hours**' vengono organizzate un giorno per l'altro e i giovani si passano di bocca in bocca, di locale in locale, il luogo stabilito per il ritrovo. (Rep)

(11) La diversificazione infatti degli orari di chiusura dei locali nel corso della notte e l'apertura dei cosiddetti **after hours** nelle prime ore del mattino, comporta un fenomeno di nomadismo fra un locale e l'altro, fra una località e l'altra, causa prima dei fenomeni così tragici di mortalità e di traumatologia. (itWaC)

(12) Per l´**after hours** il titolo è stato sospeso in attesa di comunicazioni, che l´azienda ha fornito in tardissima serata. (itWaC)

Finally, corpora are searched to verify whether Anglicisms are more, less or equally current with respect to their Italian equivalents. This information is included in each entry: in the case of *AUSTERITY* > *austerità*, *ANTI TRUST* > *antimonopolio*, *BIKINI* > *due pezzi*, the Anglicisms are more frequently used than their Italian equivalents; in the case of *ACTION MOVIE* < *film d'azione*, *AQUAPARK* < *parco acquatico*, *BISEX* < *bisessuale*, the Anglicisms are less frequent than their Italian equivalents; in the case of *ANTI-AGE* = *anti-età*, *AREA MANAGER* = *capo area*, *manager dell'area*, *ACID HOUSE* = *musica house*, the Anglicisms and their Italian equivalents are equally current.

6. Concordances

The observation of KWIC (key word in context) concordances in a corpus is extremely useful for the lexicographer to find out how a word behaves lexically and grammatically (Kilgarriff *et al* 2002). If the number of concordances is not very high, manual scanning will be enough to see how the words are used in context. For example, the search word *ABSTRACT* in the Rep produces 8 results, 5 of which correspond to the searched meaning (‘summary’), i.e. (2), (4), (5), (7) and (8):

```
1 icato come action painting, abstract expressionism, New York Scho
2 si considera, poi, che gli abstract sono quasi sempre opera di
3 da Robert Rosenblum, di "abstract sublime". Fu Harold Rose
4 di editing, redazione di abstract degli articoli e l'indi
5 ti con l’inserimento degli abstract delle notizie pubbl
6 Comincia con First italian abstract painting che nel titolo, com
7 Io, quando ho scritto l'abstract per il Congresso, ne avev
8 e allegando ad esso un "abstract" di ogni sito, di ogni pagi
```

When the number of concordances is high, manual scanning may take time and patience and sometimes be impossible, as in the case of the word *BIT* which occurs 27,000 times in the itWaC and 17,580 in the BNC. Moreover, while in Italian *BIT* is exclusively used as “a unit in computing”, in English it has a wide range of uses, including “a small piece of something” and many quantifying expressions such as “a little bit, a bit” etc. In these cases the corpus user may select various options to refine the query, filtering possible combinations of the search word or extracting its most frequent recurrent combinations. By selecting the desired combination, it is also possible to switch back and forth from the collocations to the concordance mode and visualize the whole example strings from the corpus.

For the purpose of the planned dictionary, this option has been used to find compound Anglicisms, starting from a potentially productive English word and finding its recurrent combinations. Starting from the concordance page of the word *BEAUTY,* for example, it is possible to obtain a ‘candidate list’ of collocations. This search has confirmed that *BEAUTY* combines most frequently with the items listed in Table 2, some of which are quite well-known English compounds: *BEAUTY FARM* (also in its plural form *BEAUTY FARMS*), *BEAUTY CASE*, *BEAUTY CENTER* (also occurring as *BEAUTY CENTRE*) and *BEAUTY CONTEST*.[11] These have all been included in the *Dizionario di anglicismi*.

farm	580	salon	6
case	111	queen	5
center	109	store	5
contest	31	shop	5
farms	10	fitness	4
products	9	care	3
centre	8	service	3
day	6	sleep	3
club	6		

Table 2. Collocation candidates of the word *BEAUTY* in the itWaC.[12]

7. Conclusion

This paper has illustrated the different types of lexical, grammatical and semantic information extracted from corpora in the course of the ongoing compilation of a *Dizionario di anglicismi*. Both quantitative and qualitative evidence is being used in order to decide whether potential Anglicisms are current enough to qualify as entries in the dictionary, but also to extract new anglicisms and new information that we did not have before. The observation of the data taken from two different but very large corpora has also shown that corpora, far from being all-inclusive, have various limitations imposed by what they contain (topics, registers) and the period of time that they cover. Therefore, it is up to the lexicographer to judge, for example, whether all the business and financial terms contained in the La Repubblica Corpus are actually current and representative, i.e. whether they should be recorded by a general dictionary of Anglicisms or confined to a specialised

[12] For clarity all the grammatical and hybrid combinations (with Italian words) have been excluded.

dictionary of business and financial terms. A web-based corpus like the itWaC is extremely useful to find a high number of modern and new Anglicisms in a very wide range of fields. In spite of its bias towards computer terms, videogames and music, and some 'noise' in the present version of the corpus (a good amount of 'English' is present in the Italian web pages), we may conclude that a web corpus is the best suited to the study of Anglicisms. The evaluation of both corpora, in any case, is proving to be very satisfactory and will hopefully contribute to the compilation of a well-balanced wordlist and an equally satisfactory microstructure of the forthcoming *Dizionario di anglicismi*.

References

Baroni, M., S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston and M. Mazzoleni (2004) "Introducing the La Repubblica corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper Italian", *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Vol. 5, ELRA, Lisbon, pp. 1771-1774. http://sslmitdev-online.sslmit.unibo.it/corpora/downloads/rep_lrec_ 2004.pdf

Baroni, M. and A. Kilgarriff (2006) "Large linguistically-processed web corpora for multiple languages", *Conference Companion of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, ACL, East Stroudsburg PA, pp. 87-90.

Baroni, M. and M. Ueyama (2006) "Building general- and special-purpose corpora by web crawling", *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, Tokyo, pp. 31-40.

Bauer, L. (1988) *Introducing Linguistic Morphology*, Edinburgh, Edinburgh University Press.

Crystal, D. (2003) *English as a Global Language*, Cambridge University Press, Cambridge.

De Mauro, T. and M. Mancini [2001] (2003) *Parole straniere nella lingua italiana,* Garzanti, Milano.

Furiassi, C. (2005) "Falsi anglicismi: punto d'incontro tra lessicografia e linguistica dei corpora", in G. Banti, A. Marra and E. Vineis (eds), *Atti del 4° congresso di studi dell'Associazione Italiana di Linguistica Applicata* (AItLA), Guerra Edizioni, Perugia, pp. 279-302.

Furiassi, C. and K. Hofland (2007) "The retrieval of false Anglicisms in newspaper texts", in R. Facchinetti (ed.), *Corpus Linguistics 25 Years On*, Rodopi, Amsterdam/New York, pp. 347-363.

Görlach, M. (ed.) (2001) *A Dictionary of European Anglicisms,* Oxford University Press, Oxford.

Halliday, M.A.K., W. Teubert., C. Yallop and A. Čermáková (2004) *Lexicology and Corpus Linguistics: An Introduction*, Continuum, London, New York.

Jackson, H. (2002) *Lexicography*, Routledge, London.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004) "The Sketch Engine", *Proceedings of the 11th EURALEX International Congress*, Université de Bretagne-Sud, Lorient, France, pp. 105-116.

La Repubblica Corpus. http://sslmitdev-online.sslmit.unibo.it/corpora/query

Macmillan English Dictionary for Advanced Learners (2007), Macmillan, London.

Ooi, V.B.Y. (1998) *Computer Corpus Lexicography*, Edinburgh University Press, Edinburgh.

Pulcini, V. (1999) "Focus on Italian Anglicisms: a comparative study of three dictionaries", in G. Azzaro and M. Ulrych (eds), *Transiti linguistici e culturali*. Vol. II, Edizioni Università di Trieste, pp. 359-371.

Pulcini, V. (2002) "Italian" in M. Görlach (ed.) *English in Europe*, Oxford University Press, Oxford, pp. 151-167.

Pulcini, V. (2006) "A *New Dictionary of Italian Anglicisms*: the aid of corpora", in E. Corino, C. Marello and C. Onesti (eds), *Proceedings XII EURALEX International Congress*, Vol. 1, Edizioni dell'Orso, Alessandria, pp. 313-322.

Pulcini, V. (2007) "Gli anglicismi nella lingua italiana: aspetti lessicografici", in S. Vanvolsem, S. Marzo, M. Caniato and G. Mavolo (eds), *Identità e diversità nella lingua e nella letteratura italiana. Atti del XVIII Congresso dell'A.I.S.L.L.I.* Volume primo: *L'italiano oggi e domani*, Franco Cesati Editore, Firenze, pp. 283-299.

Pulcini, V. (forthcoming) "*A Dictionary of Italian Anglicisms*: criteria of inclusion and exclusion", in G. Iamartino and N. Brownlees (eds), *Insights into English and Germanic Lexicology and Lexicography: Past and Present Perspectives*. Polimetrica, Monza.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

Sinclair, J. (2003) "Corpora for lexicography", in P. van Sterkenburg (ed.), *A Practical Guide to Lexicography*, John Benjamins, Amsterdam, pp. 167-178.

The Sketch Engine: www.sketchengine.co.uk

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*, John Benjamins, Amsterdam.

Zingarelli, N. (2006) *Lo Zingarelli 2007. Vocabolario della lingua italiana*, Zanichelli, Bologna.

3. Corpora and Translation

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 207-222 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Interjections in translated Italian: looking for traces of dubbed language

Silvia Bruti – University of Pisa
and Maria Pavesi – University of Pavia

1. Introduction

Translated language has for the past fifteen years attracted the attention of researchers coming from different fields of investigation on language and translation and mostly combining a Translation Study approach with Corpus Linguistics principles and methodology (Baker 1993, 1998; Laviosa 1998; Granger 2003; Garzone and Cardinaletti 2004; Mauranen 2005). Research in this area starts from the hypothesis that translated language has an autonomous stand and presents features which are different from both target language and source language. These specificities may derive from the social, cognitive and linguistic constraints that operate prior or during the translation process. To account for the observed regularities in both the process and product of translation, general laws, or translation universals, have consequently been postulated, which are supposed to apply to translation irrespective of the languages involved (Baker 1993; Toury 1995).

Despite the fact that many research efforts have been channelled towards identifying and describing the distinctive features of translated texts, most emphasis has been placed on the search for regular and systematic patterns in written translations, which means that little is known about the translation of ‘impromptu’ speech. In this respect, dubbing offers a good opportunity to the study of translated spontaneous spoken language (see Richet 2001), due to the naturalness pursued in film dialogues in order to ensure viewers' involvement and participation (Pavesi 2005).

Evidence of alignment with target spoken language has actually been provided for some selected grammatical features in dubbed language (e.g. Pavesi 2005, forthcoming), although unusual features with respect to spontaneous conversation have also been reported for Italian, German and Spanish audiovisual translation (e.g. Pavesi 1996; Herbst 1997; Duro 2001). Thus from a translational as well as a linguistic perspective, research should find out in which areas 'dubbese' resembles spontaneous speech and which areas, on the other hand, contribute to set it apart from the target language norms. In the latter case, on account of what can be evinced from the translation product, the influence of different universals on the translation process should be ascertained.

As highlighted by Mauranen (2004; 2005) among others, translation is a typical contact situation in which two languages are simultaneously processed by the translator (2005: 75). Such proximity and constant interchange in the translator's mind is likely to give rise to so-called interference or transfer from the source language or the source text. According to Toury (1995: 275): "in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text". Starting from the general hypothesis that interference is a translation universal and will thus affect the translation process, it is therefore worth investigating which language features are most liable or permeable to it. It is the aim of this paper to test the hypothesis that interjections – a characteristic feature of spoken language – are an area of permeability to source language influence in the translation process.

Adhering to Toury's (1980: 29-30) assumption that research on translation must start with a contrastive analysis of source and target language, we will at first provide a general overview of interjections in spoken English and spoken Italian. Corpus frequencies of the major primary interjections will then be presented for both languages. Finally, an analysis of primary interjections in a small corpus of translated films will be carried out, which will show that interjections in dubbing exhibit an atypical distribution in comparison to spoken Italian, thus contributing to the uniqueness of translated film language.

2. Interjections in English and Italian

Interjections have always been somewhat neglected in linguistic studies, either because their grammatical status has often been controversial (for a complete discussion see Wharton 2000) or because they have been collapsed with other phenomena such as exclamations (see Jovanović 2004: 19), discourse markers (see Fraser 1990; Bazzanella 1995) or inserts (Biber *et al.* 1999: 1082-1083 i.e. "stand-alone words which are characterized in general by their inability to enter into syntactic relations with other structures"). Despite this Cinderella role, however, interjections cannot be denied a clear, observable communicative power, in that they are perfectly capable of conveying the meaning of a complete utterance and offer an iconic representation of the speaker's mental or emotional attitude, or, at least, a physiological response to a percept or an object (Wharton 2000: 190-191, e.g. "Yuk! This mouthwash is foul", and "Wow! This ice cream is delicious").

The word 'interjection' derives from Latin *inter* plus *iacēre*,[1] thereby meaning "thrown in between", a meaning that well captures the minor grammatical connection interjections have with the rest of the sentence. In the literature, however, there is still debate on the nature of interjections, i.e. the 'conceptualist' approach upheld by Ameka (1992), Wierzbicka (1992) and Wilkins (1992) and the non-linguistic hypothesis advocated for example by Goffman (1981) – who regards interjections as non-linguistic,[2] i.e. "verbal gestures", "semiwords" or, more precisely, ritualised acts that are not part of language but that are to be analysed for their socio-communicative role.[3] Scholars, however, seem to agree on a general definition such as the following: interjections are self-standing units and can constitute an utterance by themselves in a non-elliptical manner and

[1] According to the *Oxford English Dictionary* (CD-ROM 1992) it probably entered the English language in the 13th or 14th century.

[2] Such a tenet is supported by several studies in aphasiology, which show that interjections are retained even in subjects who suffered severe language loss (Goodglass 1993).

[3] Wharton (2000) accommodates them on the continuum between natural, spontaneous behaviour (i.e. "showing") and linguistic, coded behaviour (i.e. "saying").

basically express a mental or emotional state.[4]

Interjections are closely tied to a specific situation in the external world, thus having a strong deictic anchorage. In most cases they are uninflected function words,[5] which due to their mobility and flexibility, can be placed almost anywhere in an utterance, without establishing syntactic dependencies with the various elements, e.g. "I am the… *ERM* new teacher", "Io sono la … *EHM*… fidanzata di Antonio" (the latter example in Poggi 1995: 408). The meaning of interjections themselves may sometimes condition their position, thereby creating some preference associations. *OH*, for example, is quite versatile both syntactically and semantically, as it can be placed in different positions with different corresponding values:

(1) *OH*, you're leaving tomorrow!

(2) The FBI arrested … *OH* … Bill Jones.
(in James 1972: 162-163)

In (1) *OH* conveys the idea that the speaker is surprised and learns something s/he did not know before, whereas in (2) s/he is having some difficulties in planning her/his utterance. On the whole, however, interjections seem to be more mobile in Italian than in English, where most often they are positioned towards the beginning of an utterance (Jovanović 2004: 21 and 27).

Interjections are traditionally distinguished into primary and secondary[6] (Poggi 1981, 1995; Ameka 1992; Jovanović 2004), depending on whether they function only as interjections or the

[4] Poggi (1995: 416) identifies three main illocutionary functions: expository, directive and phatic.

[5] But see, for example, "At the annual Dentists' convention Mrs. Pulley *WOWED* to the audience with her encyclopaedic knowledge of gold teeth" and "That is without doubt the *YUCKIEST* mouthwash I've ever tasted" in Wharton (2000: 200).

[6] See in the literature on English interjections Jovanović's (2004: 20-21) position, who distinguishes between interjections proper, short forms that are "usually one or two syllable segments" and interjections that have "originated from other parts of speech, predominantly names and adjectives, [… which] have more word-like or phrase-like forms with identifiable referents outside language or figurative meaning".

same signifier also belongs to another word class. Some of them function as complete expressions in themselves, can be defined as ‘holophrastic’ forms and have an autonomous meaning. But, more importantly, most interjections are polysemic, i.e. they can express more meanings simultaneously or the speaker can select one of the meanings according to her/his intentions.

Finally, another feature on which scholars are agreed is that interjections have graphic and phonetic traits that do not always belong to the repertoire of a language. They are mostly mono- or bi-syllabic words, which are more common in English than in Italian. Furthermore, in Italian <h> often appears either within or at the end of an interjection, in order to distinguish it from possible homographs (e.g. *AH vs A*, *EH vs E*, *OH vs O*). In both languages, differently from real words in the lexicon, vowel length and intonation often play a role in establishing meaning.

3. Translating interjections in dubbing: data from a translational corpus

In this section of this paper, we will exploit a small translational corpus[7] made up of 12 contemporary American and British films dubbed into Italian (Table 1). The corpus consists of the Italian film versions – which have been fully transcribed orthographically – for a total of about 110,000 words. The frequencies of the main Italian primary interjections will be calculated in order to evaluate to which extent dubbed Italian aligns with or differs from spontaneous spoken Italian. A more complete picture and an assessment of source language influence will also be derived from a comparison with spontaneous spoken English. To those ends, and similarly to other investigations of dubbed language (Pavesi 2005, forthcoming), corpora of spontaneous spoken language will be used as benchmarks for comparisons whose results will be used to provide a description of Italian dubbed language.

[7] A translational corpus is here to be intended as a corpus made up of translated texts, as opposed to i) a parallel corpus which contains original SL texts accompanied by their TL translations and ii) a comparable corpus, made up of translated and original texts in the same language and having the same textual features (see Granger 2003).

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

Film title	Countries of production	Year of release of Italian versions	Number of words
Ae Fond Kiss	UK/Belgium/Germany/Italy/Spain	2005	8,785
Bend it like Beckham	UK/Germany/USA	2002	9,695
Billy Elliot	UK/France	2001	5,355
Dead Man Walking	USA	1996	11,138
Crash	USA/Germany	2005	8,752
Erin Brockovich	USA	2000	12,724
Finding Forrester	USA	2001	10,582
Notting Hill	UK/USA	1999	9,635
Ocean's 11	USA/Australia	2001	9,268
One Hour Photo	USA	2002	5,252
Secrets and Lies	France/UK	1996	11,879
Sliding Doors	UK/USA	1998	8,308
Total			111,373

Table 1. The film translational corpus.

Frequency in English and Italian

Corpus-based data on interjections are available for both English and Italian. An account of frequency for English interjections is offered in Biber *et al.*'s grammar, over a sample of 1,000,000 words, distinguishing results for the two main varieties, i.e. British and American English (1999: 1096-1098). Data for Italian can be obtained from the Lessico di Frequenza dell'Italiano Parlato – LIP (De Mauro *et al.* 1993),[8] which is based on a sample of 500,000 words.

Some reflections on the frequency of English and Italian interjections are in order. As for English, Table 2 shows that *OH* is

[8] From now on referred to as LIP.

the most frequent primary interjection in both British and American varieties, followed by the interjections *Hey, Huh, Wow, Ah, Aargh, Ooh, Ha* with different orders in the two varieties.

Interjections	**American English**	**British English**
Oh	4,000	4,000
Hey	300	50
Huh	250	25
Wow	200	25
Ah	150	550
Aargh	150	25
Ooh	100	350
Ha	50	200
Ops	25	<12,5
Whoops	25	<12,5
Aha	<12,5	100
Whoa	<12,5	25
Tt	<12,5	25

Table 2. Frequency of interjections in English according to Biber *et al.* (1999) normalized to 500,000 words.

Data for Italian (De Mauro *et al.* 1993) are detailed in Table 3 below. As can be seen, the most frequent interjection in Italian is *Eh/Eh?,*[9] with 7,542 occurrences, followed by *Ah, Ahah, Be'/Be'?, Oh.* The other interjections occur less frequently and include many items that are socially and geographically marked such as *Aho'/Ohe'/Ola'*.

[9] *Eh* has two main but distinct functions in statements and in questions. The same distinction applies to *Be'*.

EH/EH?	7,542	***SS***	10
AH	2,174	***EHI***	9
AHAH	1,205	***M***	5
BE'/BE'?	325	***TZ***	5
OH	79	***BLEAH***	4
MAH	71	***AHIME'***	1
BO'	70	***UH***	11
AHO'	67	***ALE'***	1
EMBE'/EBBE'/'MBE'/MBE'	37	***EHM***	1
UHE'	27	***NE'***	1
IH	26	***OLA'***	1
OHE'	19	***TO'***	1
BAH	13	***TSK***	1
AHI (6)/ OGLIOHOI (4)/ AGLIA (1)	11	**Total**	**11,717**

Table 3. Frequency of interjections in Italian in the LIP corpus (1993).

In both English and Italian only few interjections occur extremely often, thus covering the most part of the use of interjections in spoken language as shown in Table 4 below, where frequencies of the most common ones in English and in Italian are reported for comparison.

Italian		**American English**		**British English**	
EH	7,542	***OH***	4,000	***OH***	4,000
AH	2,174	***HEY***	300	***AH***	550
OH	79	***HUH***	250	***HEY***	50
EHI	9	***AH***	150	***HUH***	25

Table 4. Frequency orders of some of the most frequent interjections normalized to 500,000 words.

Frequencies in dubbed Italian

In order to proceed to a comparison with spoken Italian, the frequencies of the main Italian primary interjections have been extracted from the translational corpus[10] (Table 5) and they have been compared with the frequency list from the LIP corpus.

Interjections[11]	**Raw frequencies**	**Normalized frequencies (to 500,000)**
Ah	511	2,294
Oh	412	1,850
Eh/Eh?	284	1,275
Be'/Be'?	309	1,387
Ehi	111	498
Ahah	7	31
Wow	3	13
Aho'/Ohe'/Ola'	0	0
Mah	4	18
Bo'	0	0
Embe'/Ebbe'/Mbe'/Mbe' ?	0	0
Uhe'	0	0
Ih	0	0
Bah	0	0
Ahi (6)/Ogliohoi (0)/Aglia (0)	6	30
Uh	28*	126
Ss	4	18
M	0	0
Tz	0	0
Beah	1	4
Aime'	1	4
Ae'	0	0
N	0	0
T	0	0
Tsk	0	0
	1,681	**7,548**

Table 5. Frequencies of interjections in the film translational corpus.[12]

[10] With the software WordSmith Tools (Scott 1996).

[11] For the sake of uniformity, we have adopted the orthographic conventions used in the LIP corpus, i.e. apostrophes and not accents.

A first comparison shows that the repertoire of primary interjections in dubbed Italian is much reduced: out of the 34 interjections which occur in the LIP corpus, only 14 appear in the translated films, with the addition of *WOW* [waʊ] bringing the number to 15. A smaller repertoire of given items diminishes the expressive potential of translated language, while at the same time producing an inevitable effect of repetitiveness and automaticity. The formulaic nature of dubbed language in general has in fact been pointed out for Italian translated from English as well as for other dubbed languages such as Spanish and Catalan (see Chaume's (2001) notion of 'pre-fabricated orality').

Reduction, however, does not only affect types but extends to tokens as well. Overall, interjections are less frequent in the corpus of translated Italian than in the LIP corpus, with interjections in dubbed Italian amounting to only 64% of those in spoken Italian (7,548 *vs* 11,711). This result aligns with the already observed use of fewer 'peripheral' features such as weak connectors (*POI*, *CIOÈ*, *ALLORA*), vocatives and swear words in translated film language as opposed to both target language and source texts (Pavesi 2005). Furthermore, the difficulty to describe the meaning and the use of primary interjections (Richet 2001) together with their strong deictic anchorage may contribute to the observed avoidance phenomena which produce reductions in the repertoire and in the occurrences of 'minor' interjections in dubbed language.

The most interesting trend extracted from the translational corpus, however, has to do with the frequency of individual primary interjections, which crucially differs from their frequency in the target language. Although on the whole the most frequent interjections in spoken Italian are also frequent in dubbed Italian, a comparison of the frequency orders from the whole LIP and the translational corpus shows that spoken Italian and translated film language do not exhibit the same patterns. Some interjections are under-represented in dubbese and others are greatly over-represented. The most frequent Italian interjections *EH* and *EH*?, which together occur 7,542 times in the LIP and make up for much more than half of all interjections in spoken Italian, occur only 284

[12] An asterisk * indicates that interjections repeated more than twice in a row have not been counted.

times in the translational corpus (normalized frequency = 1,275). This means they are almost 6 times less frequent in dubbed language.

Conspicuous drops in frequencies also occur for less frequent although very typical Italian interjections that do not have a phonic equivalent in English, such as *MAH*, *BLEAH* and *BO'*. Some other specific Italian interjections appear to be the first to be excluded from dubbed Italian: these are *NE', TO'*, *ALE'*. Among other excluded interjections we find those which are geographically or socially marked, such as *EMBE', EBBE', MBE'*, and *AHO'*, *UHE'*, *OHE'*, *OLA'*. These frequency patterns are in line with the universal of standardization which predicts the preference in translations for the most central, prototypical items of the target language.

If many Italian interjections appear to be under-represented in dubbed Italian, others are considerably over-represented in the corpus of translated film dialogues. In particular *OH* with its 412 tokens is 23 times as frequent in dubbese as in spoken Italian. *EHI*, which occurs 111 times in the translational corpus, appears to be 55 times as frequent in dubbed Italian as in Italian speech. *BE'* occurs 309 times in the translated films, which means it is 4 times as frequent as in the LIP. Finally, 28 tokens of *UH* in the translation corpus represent the interjection 10 times more often than it occurs in the target language.

These frequencies in themselves set dubbed Italian apart from spontaneous spoken language and identify interjections as an area where deviations from the target language occur, thus empirically contributing to the characterization of Italian dubbed language as a third code (see Pavesi 1996, but also Maraschio 1982; Raffaelli 1996; Antonini and Chiaro forthcoming).

Interference in translation

But what makes of interjections an area of differentiation between dubbed Italian and Italian target language? Why are some interjections under-represented and others over-represented? As pointed out earlier, some of the deviations from target language use may be induced by the polysemy, semantic and pragmatic looseness of interjections both in the source and the target language. Uncertainty about the meaning and the function of the interjections, but also their traditionally recognized interchangeability (Richet 2001), may in fact cause greater transfer of the source language item during the translation process. Transfer from English into Italian can also be induced by the phonetic similarity between some English and Italian items, dubbing being a translation context which favours phonetically close equivalents.

More specifically, we want to argue that the Unique Items Hypothesis as postulated by Tirkkonen-Condit (2002, 2004), Eskola (2004), Mauranen (2005) contributes to account for the frequency patterns of major primary interjections in dubbed Italian. According to the hypothesis

> translations tend to under-represent target-language specific, unique linguistic features and over-represent features that have straightforward translation equivalents (functioning as some kind of stimuli) in the source language (Eskola 2004: 83).

The bias towards items which share features between source and target language may explain the higher than expected frequencies of *OH*, *EHI* and *UH* in dubbed Italian as opposed to original spoken language. In particular, English *OH*, the most frequent interjection in both American and British English, is phonetically very similar to the Italian *OH*. The English interjection also shares some of its semantic and pragmatic functions with the Italian equivalent, with *OH* conveying surprise in front of unexpected information or feelings of disappointment and sorrow in both languages. The semantic and pragmatic overlap may well represent a further motivation for transfer from source texts to target texts.

Along similar lines, it can be argued that translators introduce *EHI* into Italian dubbing with a frequency higher than expected in the target language due to similarities in the phonetic realizations and the functions of *HEY/EHI* in the two languages. Both interjections are used as a call for attention, especially when the speaker means to address someone in particular. English *HEY*, however, is much more frequent, especially in American English, than the Italian *EHI*.

As for *UH*, the graphemic similarity between the two interjections in the two languages (pronounced [ʌ] in English *vs* [u:] in Italian) justifies a degree of over-representation of the interjection in dubbed

Italian, despite the fact that the two interjections differ functionally in the two languages.

Finally, although no direct data is available at present (Rolandi 2005), the over-representation of *BE'* in dubbed Italian is also likely to be source-language related. *BE'* is in fact one of the most frequent translations of the English discourse marker *WELL*, typically used as an utterance launcher. Similarly to *WELL*, *BE'* derives from *bene* (adv., 'well'), of which it is a contraction. Its function is that of expressing hesitation on the part of the speaker and distancing from what has just been said by the interlocutor.

In this context of target language influence on translation outcomes, it does not appear to be by chance that the most frequent interjection in spoken Italian, *EH*, is given relatively low emphasis in dubbed language. There is actually no phonetically matching interjection in English, nor a straightforward functional equivalent.[13] Italian *EH* has a plurality of semantic and pragmatic meanings, such as confirmation or request for confirmation or action as well as agreement and elicitation (also with oneself). Lacking both phonetic and semantico-pragmatic correspondence with English, *EH* is unique to Italian and therefore less susceptible to be used in translated texts. In line with the Unique Items Hypothesis (Tirkkonen-Condit 2002, 2004; Eskola 2004; Mauranen 2005) there is a bias in translation against items which are specific to the target language.

4. Conclusion

The present study has shown that interjections in dubbed Italian are globally less frequent and less varied than in spontaneous speech. More interestingly, the frequency patterns in the film translation corpus differ from the corresponding patterns in the LIP corpus, thus highlighting the self-standing status of dubbed Italian. Moreover, there is strong evidence of the influence of English on these 'marginal' features of this variety of translated Italian. Italian interjections which exhibit a degree of similarity with English ones

[13] If we exclude the interrogative interjection *EH?*, restricted to some social and geographical varieties of English.

tend to be over-represented in dubbing, whereas interjections which are specific and restricted to Italian tend to be under-represented.

The data presented here thus offer further support to the Unique Items Hypothesis. The hypothesis is clearly interference-based and has the advantage of accounting for many reductions or simplifications in translations. However, why precisely some language areas are subjected to interference, while others pattern more or less independently of the source language still begs the question and deserves further research in investigations of translated language.

References

Antonini, R. and D. Chiaro (forthcoming) "The general perception of likelihood of occurrence: the translational norms of language-specific features on the Italian screen", in G. Anderman and J. Díaz Cintas (eds), *In So Many Words: Language Transfer on the Screen*, Multilingual Matters, Clevedon.

Ameka, F. (1992) "Interjections: the universal yet neglected part of speech", *Journal of Pragmatics* 18 (2/3), pp. 101-118.

Baker, M. (1993) "Corpus linguistics and translation studies: implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam/ Philadelphia, pp. 233-250.

Baker, M. (ed.) (1998) *Encyclopaedia of Translation Studies*, Routledge, London/ New York.

Bazzanella, C. (1995) "I segnali discorsivi", in L. Renzi, G. Salvi and A. Cardinaletti (eds), *La grande grammatica italiana di consultazione*, Vol. III, Il Mulino, Bologna, pp. 225-257.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *The Longman Grammar of Spoken and Written English*, Longman, Harlow.

Garzone, G. and A. Cardinaletti (eds) (2004) *Lingua, mediazione linguistica e interferenza*, Franco Angeli, Milano.

Chaume, F. (2001) "La pretendida oralidad de los textos audiovisuales y sus implicaciones en traducción", in F. Chaume and R. Agost (eds), *La traducción en los medios audiovisuals*, Publicacions de la Universitat Jaume I, Castelló, pp. 77-87.

De Mauro, T., F. Mancini, M. Vedovelli and M. Voghera (1993) *Lessico di frequenza dell'italiano parlato*, Etas Libri, Milano.

Duro, M. (2001) "'Eres patético': el español traducido del cine y de la televisión", in M. Duro (ed.), *La traducción para el doblaje y la subtitulación*, Cátedra, Madrid, pp. 161-351.

Eskola, S. (2004) “Untypical frequencies in translated language: a corpus-based study on a literary corpus of translated and non-transalted Finnish”, in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp.83-99.

Fraser, B. (1990) “An approach to discourse markers”, *Journal of Pragmatics* 14 (3), pp. 383-395.

Goffman, E. (1981) *Forms of Talk*, Blackwell, Oxford.

Goodglass, H. (1993) *Understanding Aphasia*, Academic Press, New York.

Granger, S. (2003) “The corpus approach: a common way forward for contrastive linguistics and translation studies”, in S. Granger, J. Lerot and S. Petch-Tyson S. (eds), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Rodopi, Amsterdam/Atlanta, pp. 17-29.

Herbst, T. (1997) “Dubbing and the dubbed text – style and cohesion”, in A. Trosborg (ed.), *Text Typology and Translation*, John Benjamins, Amsterdam/ Philadelphia, pp. 291-308.

Jovanović, V.Ž. (2004) “The form, position and meaning of interjections in English”, *Facta Universitatis. Series: Linguistics and Literature* 3 (1), pp. 17-28.

Laviosa, S. (1998) “Universals of translation”, in M. Baker (ed.), *Encyclopaedia of Translation Studies*, Routledge, London/New York. pp. 288-291.

Maraschio, N. (1982) “L’italiano del doppiaggio”, in AAVV, *La lingua italiana in movimento*, Accademia della Crusca, Firenze, pp. 135-58.

Mauranen, A. (2004) “Corpora, universals and interference”, in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 65-82.

Mauranen, A. (2005) “Contrasting languages and varieties with translational corpora”, *Languages in Contrast* 5 (1), pp. 73-92.

Mauranen A. and P. Kujamäki (eds) (2004) *Translation Universals. Do they exist?*, John Benjamins, Amsterdam/Philadelphia.

Oxford English Dictionary on CD-ROM (1992).

Pavesi, M. (1996) “L’allocuzione nel doppiaggio dall’inglese all’italiano”, in C. Heiss and R.M. Bollettieri Bosinelli (eds), *Traduzione multimediale per il cinema, la televisione e la scena. Atti del convegno internazionale “Traduzione multimediale per il cinema, la televisione e la scena, Multimediale Übersetzung für Film, Fernsehen und Bühne, Multimedia Translation for Film, Television and the Stage”*, Cooperativa Libraria Universitaria Editrice, Bologna, pp. 117-130.

Pavesi, M. (2005) *La traduzione filmica. Aspetti del parlato doppiato dall’inglese all’italiano*, Carocci, Roma.

Pavesi, M. (forthcoming) “Spoken language in film dubbing: target language norms, interference and translational routines”. In D. Chiaro and C. Heiss (eds), *Between Text and Image. Updating Research in Screen Translation*, John Benjamins, Amsterdam/Philadelphia.

Poggi, I. (1981) *Le interiezioni: studio del linguaggio e analisi della mente*, Boringhieri, Torino.

Poggi, I. (1995) “Le interiezioni”, in L. Renzi, G. Salvi and A. Cardinaletti (eds), *La grande grammatica italiana di consultazione*, Vol. III, Il Mulino, Bologna, pp. 403-427.

Raffaelli, S. (1996) “Un italiano per tutte le stagioni”, in E. Di Fortunato and M. Paolinelli (eds), *Barriere linguistiche e circolazione delle opere audiovisive: la questione del doppiaggio*, AIDAC, Roma, pp. 25-28.

Richet, B. (2001) “Quelques données et réflexions sur la traduction des interjections”, in M. Ballard (ed.), *Oralité et traduction*, Presses Université, Artois, pp. 79-127.

Rolandi, C.C. (2005) *La traduzione di interiezioni primarie nel film* Out of sight. *Opposites attract – Gli opposti si attraggono*, unpublished MA thesis, University of Pavia.

Scott, M. (1996) *WordSmith Tools Manual*, Oxford University Press, Oxford.

Tirkkonen-Condit, S. (2002) “Translationese – a myth or an empirical fact? A study into the linguistic identifiability of translated language”, *Target* 14 (2), pp. 207-220.

Tirkkonen-Condit, S. (2004) “Unique item – over- or under-represented in translated language?”, in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 177-184.

Toury, G. (1980) “Contrastive linguistics and translation studies: towards a tripartite model”, in G. Toury (ed.), *In Search of a Theory of Translation*, The Porter Institute for Poetics and Semiotics, Tel Aviv, pp. 19-34.

Toury, G. (1995) *Descriptive Translation Studies and Beyond*, John Benjamins, Amsterdam/Philadelphia.

Wharton, T. (2000) “Interjections, language and the ‘showing’/’saying’ continuum”, *UCLWPL (Phonetic and Linguistics)* 12, pp. 173- 215.

Wierzbicka, A. (1992) “The semantics of interjection”, *Journal of Pragmatics* 18 (2/3), pp. 159-192.

Wilkins, D.P. (1992) “Interjections as deictics”, *Journal of Pragmatics* 18 (2/3), pp. 119-158.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 223-238 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Corpus studies of translation universals: a critical appraisal

Sara Laviosa – University of Bari

1. Introduction

Divergently similar to Baker, is Toury's view on universals. He regards them as conditioned and probabilistic regularities in translation, and prefers the term 'laws' rather than 'universals' mainly because "it should always be possible to explain away [seeming] exceptions to a law with the help of *another* law, operating on *another* level" (Toury 2004: 29). The value of such probabilistic laws of translational behaviour, he argues, lies on their "*explanatory power*" rather than their "*existence*" (Toury 2004: 29). Toury puts forward two exemplary laws of translational behaviour: the law of growing standardisation and the law of interference. In their general form, without specifying any conditioning factors, the two laws read as follows:

> in translation, textual relations obtaining in the original are often modified, sometimes to the point of being totally ignored, in favour of [more] habitual options offered by a target repertoire (Toury 1995: 268);
>
> in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text. (Toury 1995: 275).

In a similar vein, Chesterman (2000, 2004) views the quest for universal features of translations as one way in which descriptive scholars propose and look for generalizations about translation. These general regularities or laws, he explains, are explored by putting forward, operationalizing and testing, through a comparative model of translation, general descriptive hypotheses about the existence of similarities between different types of translation, without disregarding either the differences between them or the uniqueness of each particular case. Chesterman (2004: 39) makes a useful distinction between S-universals, which refer to “universal differences between translations and their source texts”, and T-universals, which refer to “universal differences between translations and comparable non-translated texts”. If universals, which are essentially descriptive constructs, are supported by extensive empirical evidence, they can have explanatory force as regards the occurrence of a given feature in a particular translation (Chesterman 2000: 26). The causes of universals, on the other hand, are to be found not only in the nature of translation as a communicative act and the translator’s awareness of his/her socio-cultural role, but also in neighbouring fields of scientific enquiry, such as human cognition. Therefore in Translation Studies, like in any discipline, general explanatory laws not only permit to make predictions about future cases but also create vital interdisciplinary links (Chesterman 2000, 2004).

Drawing on Croft’s (1990: 246) “scalar concept of generalization”, Halverson (2003: 232) posits that universals are second-level (or internal) generalizations made over numerous empirical studies, and as such they are “*explanatory* with respect to individual studies of particular linguistic realizations and/or language pairs”. On the other hand, third-level (or external) generalizations are made on the basis of cognitive factors. From this psycholinguistic perspective, Halverson argues that various universal lexical/semantic patterns observed in ST-TT pairs, parallel corpora and monolingual comparable corpora can be explained by the existence of asymmetries in the cognitive organization of semantic information, whereby the nodes which function as category prototype and highest-level schema are more prominent and important than others, mostly as a result of their high frequency of use (Langacker 1987). Conversely, the absence of these asymmetries is assumed to produce the opposite effect in translated text.

The notion of translation universals has been the object of lively debate. Tymoczko (1998: 653), for example, contends that the search for universals is firmly grounded in empiricism, whose claims about scientific objectivity have been seriously challenged by modern thought. Similarly, in reply to Chesterman's claim that "a statement about a regularity is first of all a descriptive hypothesis" (Chesterman and Arrojo 2000: 157), Arrojo affirms that the regularities that may be unveiled in translation "will reflect the interests of a certain translation specialist, or a research group, at a certain time, in a certain context" (Chesterman and Arrojo 2000: 159). She therefore rejects the essentialism allegedly inherent in the concept of universal. Baker (2001) adds her voice to the current debate, expressing her doubts about the appropriateness of the choice of the term 'universals' to refer to typical patterns of translational behaviour. However controversial the notion of universals may be, there is no doubt that the introduction of electronic corpora in Translation Studies has acted as a stimulus to empirical research into this aspect of translational language, as well attested by a recent collected volume on the subject, edited by Anna Mauranen and Pekka Kujamäki (2004), Maeve Olohan's (2004) book on corpora and Translation Studies, as well as two international conferences: *Corpus-based Translation Studies: Research and Applications*, 22 - 25 July 2003, Pretoria and *Conference and Workshop on Corpora and Translation Studies*, 30 March–1 April 2007, Shanghai.

2. Empirical evidence and counter evidence

Thanks to the availability of parallel and comparable corpora in a growing number of languages, corpus-based studies have refined and diversified previous descriptive research into several translation universals, most notably simplification, explicitation and normalization. What follows is a critical review of some of the main studies carried out in this expanding area of scholarly enquiry.

Simplification

Four "core patterns of lexical use" were identified by Laviosa in the English Comparable Corpus (ECC), a multi-source-language monolingual comparable corpus made up of translational and non-translational narrative and newspaper texts:

- translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower);
- the proportion of high frequency words versus lower frequency words is relatively higher in translated texts;
- the list head of a corpus of translated text accounts for a larger area of the corpus (i.e. the most frequent words are repeated more often);
- the list head of translated texts contains fewer lemmas. (Laviosa 1998: 565).

Largely independently of the influence of the source language, translators seem to restrict the range of words available to them and use a relatively higher proportion of high-frequency lexical items. The findings confirm the general hypothesis of lexical simplification in translation, defined by Blum-Kulka and Levenston (1983: 119) as “the process and/or result of making do with *less* words”. The above patterns of lexical use, argues Halverson (2003: 218-219), support and can be accounted for by the idea of gravitational pull from category prototypes in semantic networks, since prototypes are selected more frequently than more peripheral structures or items.

Explicitation

Ǿverås (1998) tested Blum-Kulka’s general Explicitation Hypothesis in a corpus of literary translations drawn from the bidirectional English-Norwegian Parallel Corpus (ENPC). The explicitation hypothesis “postulates an observed cohesive explicitness from SL to TL texts regardless of the increase traceable to differences between the linguistic and textual systems involved” (Blum-Kulka 1986: 19). Assuming that a rise in the level of cohesion in the target language text is an aspect of explicitation, Ǿverås hypothesized that English and Norwegian target texts would be more cohesive than their source texts. The results largely confirmed this prediction since the explicitating shifts, involving the addition and specification of lexical and grammatical items, were found to outnumber the implicitating shifts in both direction of translation, although English target texts showed a lower level of explicitness vis-à-vis Norwegian target texts. In addition to the process of interpretation inherent in translation, Ǿverås considers various factors that may explain the phenomenon of explicitation, for example the stylistic preferences of the source and target languages, their systemic differences, and culture-bound translation norms.

Baker's (1996: 180) notion of explicitation, which refers to "an overall tendency to spell things out rather than leave them implicit in translation", was investigated by Olohan and Baker (2000) at the level of syntax through the analysis of the occurrences of the reporting *THAT* in translated fiction and biography texts drawn from the Translational English Corpus (TEC) vis-à-vis comparable originals drawn from the British National Corpus (BNC). The findings show a preference for the use of the optional *THAT* with the verbs *SAY* and *TELL*, and suggest a higher level of grammatical explicitness in translational English. Drawing on Rohdenburg (1996), the authors suggest that the cognitive complexity involved in translation can explain the over-representation of the optional *THAT* in translated texts. Complementary findings were obtained by Kenny (2005), who found that in the unidirectional German-English Parallel Corpus of literary texts (GEPCOLT), patterns of omission of the optional *THAT* in English tended to reflect patterns of omission of the optional *DAß* in German, whereas patterns of inclusion of the optional *THAT* did not reflect patterns of inclusion of *DAß* (Kenny 2005: 160). In addition, Olohan's (2003, 2004) recent studies reveal a correlation in both translated and non-translated texts between omission of the optional *THAT* and use of contracted forms, so that, while TEC texts are more likely to include *THAT* and not use contractions, BNC texts are more likely to omit *THAT* and use contractions. Moreover, TEC texts appear to contain a more standard variant of the English language, with fewer dialectal or sociolectal markers (Olohan 2004: 104).

Drawing on Blum-Kulka's explicitation hypothesis and Baker's notion of explicitation, Pápai (2004) put forward three specific hypotheses which she examined with the ARRABONA corpus, a combined English-Hungarian parallel corpus and a corpus of comparable original Hungarian texts. Pápai predicted first of all that

English-Hungarian translations would be characterized by five explicitation strategies involving not only shifts in cohesion, but also addition of linguistic and extra-linguistic information and disambiguation of ST items. The hypothesized strategies were: 1) addition and modification of punctuation marks, 2) addition of derivatives, 3) addition of conjunctions, 4) addition of conjunctions and cataphoric reference, 5) addition of discourse particles. It was also predicted that translated Hungarian texts would show a higher level of explicitness than comparable originals, and the degree of explicitness would be higher in translated scientific texts than in literary texts. The first two hypotheses were confirmed, the third one was not supported.

Normalization

Normalization, defined by Baker (1996: 176-177) as “the tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them”, was the starting point of Kenny’s (2001) study of lexical creativity and lexical normalization in a parallel corpus of contemporary German literary texts and their English translations (GEPCOLT). Kenny identified three sets of creative lexis in the German subcorpus: creative word forms identified from an initial list of *hapax legomena* (word forms that occur only once in the corpus), creative forms specific to a particular writer, and creative author-specific collocations. She found that 44% of creative *hapax legomena* and 16% of creative collocations were normalized. So, although lexical normalization was found to be a feature of translation, on most occasions normalization did not take place. The extent to which creative lexis is normalized appears to be influenced by how the translator sees his/her brief and on the systemic resources of the source language, for example creative lexis linked to the derivational possibilities offered by German, or those that involve puns might be particularly difficult to render in the target language. Some evidence of normalization is also provided by Ǿverås’ (1998) study of explicitation discussed earlier, which shows a tendency in translation to prefer typical rather than unusual collocations and neutralize metaphorical expressions.

The findings obtained in these two studies lend some support to Toury’s law of growing standardisation. In the process of translation, argues Toury, the dissolution of the original set of textual relations is inevitable and can never be fully recreated. Moreover, he suggests that factors such as age, extent of bilingualism, the knowledge and experience of the translator, as well as the status of translation within the target culture may influence the operation of the law (Toury 1995).

On the other hand, Saldanha's (2004) hypothesis that the use of split infinitives would be less common in translational versus non-translational English narrative, consistently with the general hypothesis of standardisation in translated language, was not confirmed in a study of a monolingual comparable corpus made up of a subset of the Translational English Corpus (TEC) and a subset of the British National Corpus (BNC). In an attempt to account for this exception to normalisation, Saldanha (2004: 49) suggests that "the use of split infinitives could be an element of the translator's style, one of the linguistic habits that characterise the work of some translators and distinguishes it from others".

Discourse transfer

Discourse transfer refers to the translators' tendency to produce a translated utterance not by retrieving the target language via their own linguistic knowledge, but directly from the source utterance itself (Toury 1986). The universality of discourse transfer is expressed by Toury's law of interference, whereby the transfer of source-text phenomena can be of two types: "*negative transfer* (i.e. deviations from normal, codified practices of the target system" and "*positive transfer* (i.e. greater likelihood of selecting features which do exist and are used in any case" (Toury 1995: 275). The operation of the law of interference depends on the particular manner in which the source text is processed, so that "the more the make-up of a text is taken as a factor in the formulation of its translation, the more the target text can be expected to show traces of interference" (Toury 1995: 276). The extent to which interference is actually realized depends also on the professional experience of the translator and on the socio-cultural conditions in which a translation is produced and consumed. These two factors are built into the law of interference as conditions, so that

> even when taking the source text as a crucial factor in the formulation of its translation, accomplished translators would be less affected by its actual make-up. (Toury 1995: 277)

and

> tolerance of interference – and hence the endurance of its manifestations - tend to increase when translation is carried out from a ‘major’ or highly prestigious language/culture, especially if the target language/culture is ‘minor’, or ‘weak’ in any other sense. (Toury 1995: 278).

The degree of tolerance towards interference is affected not only by the relationship of dominance and prestige underlying two language communities, but also by the prestige value attached to different text types, so that there may well be, within one target culture and with respect to the same source culture, less tolerance towards interference in technical translation than in literary translation (Toury 1995: 278-279).

In her investigation of multi-word strings that convey text-reflexive meanings in the Corpus of Translated Finnish (CTF), Mauranen (2000) draws on Toury’s law of interference and Robinson’s (1997) claim that translations from a highly prestigious into a less prestigious culture which are aimed at a popular readership tend to be written in a fluent style. Mauranen put forward two hypotheses. The first was that as a result of the law of interference, Finnish academic texts translated from English vis-à-vis comparable original Finnish texts would show a higher frequency of multi-words strings with the function of organising the text, providing comments and guiding the reader’s interpretation. The second hypothesis was that the effect of the law of interference would be less noticeable in popular non-fiction. The first hypothesis was confirmed. Moreover, translated academic texts showed a different and/or more varied pattern of near synonymous lexical combinations compared with original texts, and the highly target language specific item *TOISAALTA*, which roughly means *ON THE OTHER HAND*, but has no exact equivalent in English, was hugely under-represented in translated texts, independently of source language and genre variation. Both the over-representation of text-reflexive expressions in translational texts versus comparable originals and the under-representation of TL-specific lexical items can be regarded as examples of negative discourse transfer. More precisely, the former transfer is an example of what Mauranen calls “pair-wise interference”, that is interference which is specific to a particular language pair, while the latter exemplifies the law of interference as a “universal language-independent law” (Mauranen 2004: 69). The second hypothesis was not confirmed. Translations of the lower-prestige genre of popular non-fiction deviated more from target language norms than the translations of the higher-prestige genre of academic prose. Possible reasons for this difference may be the professional status of the translator of academic texts, who may devote more time to the translation task, and the critical evaluation of the linguistic qualities of translated academic writing (Mauranen 2000: 137).

Nilsson (2004) compared the frequency and collocational patterns of the Swedish function word *AV* (‘of’, ‘by’) in narrative texts drawn from the English-Swedish Parallel Corpus (ESPC), a bidirectional parallel corpus of English and Swedish fiction and non-fiction. Her findings showed a significantly higher frequency of *AV* and its typical collocational and colligational patterns in translated Swedish vis-à-vis comparable originals, both results being attributed mainly to the influence of the source language.

In a study carried out with an English-Italian parallel corpus of biology university books, Pavesi (2003) showed that motion events as they are realized in the combination of a verb of motion plus the preposition *INTO* were rendered in Italian by equivalent expressions characterised by:

- loss of Manner expressed by the motion of verb in English;
- loss of detail in the description of complex Path;
- loss of the meaning of boundary crossing with non-telic verbs of motion. (Pavesi 2003: 155-156).

These lexico-semantic patterns were not as frequent as expected on the basis of structural and rhetorical differences between English and Italian, possibly as a result of the translator’s attempt to transfer as much information as possible in a type of specialized text with a predominant didactic communicative function. This kind of under-representation of a target language structure in translated texts can be regarded as a form of negative transfer.

Further evidence of the operation of the law of interference is provided by Musacchio's (2005) investigation of Anglicisms, carried out with an English-Italian parallel corpus and an Italian comparable corpus of original newspaper and magazine articles on business and economics. Her findings showed several examples of negative transfer at the level of syntax, where translations exhibited close renderings of syntactic constructs and higher frequency of possessive determiners, demonstrative determiners and demonstrative pronouns.

Tirkkonen-Condit's proposed Unique Items Hypothesis (UIH), which can be subsumed under either Toury's general law of interference as a particular case of negative discourse transfer or Baker's (1993: 245) universal known as the distinctive distribution of target-language items, states that target language specific elements, which do not have equivalents in the source language, tend to be under-represented in translated texts compared with comparable originals, since "they do not readily suggest themselves as translation equivalents" (Tirkkonen-Condit 2004: 177-178). The hypothesis was tested on two subcorpora drawn from the Corpus of Translated Finnish: academic and fictional texts. Two sets of elements specific to Finnish were investigated: verbs of sufficiency and the clitic particles *–KIN* and *–HAN*. The findings strongly supported the hypothesis notwithstanding some differences between the two genres. As discussed earlier, Mauranen (2000) too found confirmation of the Unique Items Hypothesis when examining the occurrence of a TL-specific item, *TOISAALTA*, in translated and non-translated academic writing and popular non-fiction.

3. Universals in the translation classroom

Translation universals have recently found a place also in the applied extensions of the discipline, where they have been usefully employed in the subfields of translator training and translation quality assessment. Normalization, for example, was the point of departure of Stewart's (2000) corpus study of conventionality in the context of teaching L2 translation. Stewart's classroom-based research into the use of the British National Corpus (BNC) for translating tourist brochures from Italian into English as L2 showed that translator trainees could produce naturally sounding collocations by examining the frequency of occurrence and concordance lines of assumed target language equivalents of source language noun phrases. Two examples of corpus-based translation equivalents were *GRAN GIRO DELLA CITTÀ* and *GRAND TOUR OF THE CITY*; *STRADA PANORAMICA* and *ROAD WITH PANORAMIC VIEWS*. A large corpus such as the BNC can therefore be a very useful resource for students translating into English as a foreign language since it can compensate for their lack of native-speaker knowledge of target language and culture. However, argues Stewart, the use of corpora in the translation classroom may well contribute to reinforcing the normalizing tendency displayed by translated texts and maybe inhibit creativity.

The Unique Item Hypothesis was tested experimentally in the translation classroom by Kujamäki (2004) with a view to raising student awareness of what the translation process entails. In the first phase of the experiment, thirty-six students were asked to back-translate into Finnish the German and English translations of a Finnish original text created *ad hoc* on the topic of driving in Finland, which included several language-specific items with no straightforward equivalents in either German or English. In the second phase of the experimental design the students' translations were compared with the students' use of original Finnish as revealed by a cloze test designed to elicit 'unique items'. The findings confirmed the UIH hypothesis. In their translations, in fact, students tended to overlook unique items and opt for straightforward lexical or dictionary equivalents, even when TL-specific items are part of their lexical repertoire, as revealed by the results of the cloze test.

Finally, within the research area of Corpus-based Translation Quality Assessment (TQA), Scarpa (2006) assessed, in relation to specialized translations carried out by advanced translator trainees, the validity of simplification and explicitation as translation universals and possible indicators of translation quality. The aim of the study was to deepen our understanding of the nature of translation with a view to providing a firm basis for evaluation. Specialized English-Italian translations carried out by advanced translator trainees were first compared with the English source texts as regards overall length, number of sentences, average sentence length, standardized type/token ratio, and lexical density (calculated as the ratio of grammar to content words). The results of this initial analysis were then compared with the TQA grades given by the evaluators. In the last stage, the translated texts were compared with comparable originals drawn from the Italian reference corpus CORIS (Corpus dell'Italiano Scritto). The Italian translations were generally found to be longer and have fewer and longer sentences than the English originals, standardized type/token ratio was higher and lexical density lower. Compared to lower-scoring translations, the higher-scoring ones had a lower number of running words, higher average sentence length, lower number of sentences, higher type/token ratio, and lower lexical density. Higher- and lower-scoring translations deviated from comparable originals on all these measures. Simplification and explicitation were largely confirmed by comparisons made with source texts and comparable originals. The analysis of the relation between universals and TQA grades showed that higher-scoring translations had a higher level of syntactic explicitness and a lower level of lexical simplification compared with lower-scoring translations.

4. Conclusion

The quest for translation universals is pursuing one research aim and is giving rise to a variety of investigations that are divergently similar as regards their rationale, the object of study, the research model and methodology adopted, the findings, and the posited reasons for the existence of universals. The importance of this variegated area of research has been recognised by many scholars. Chesterman (2004: 46), for example, claims that the study of universals

> has been one of the most important methodological advances in Translation Studies during the past decade or so, in that it has encouraged researchers to adopt standard scientific methods of hypothesis generation and testing.

Hermans (2006: 86), on the other hand, contends that the search for universals

> is compromised by the fact that the available translation corpora cover only a limited number of languages, lack a historical dimension and have no way of identifying whether the features encountered are exclusive to translation.

There is no doubt, in my view, that corpus studies of universals have pushed the discipline towards empiricism and helped unveil the linguistic features that characterise translation as a language variety in its own right. What can be envisioned for the bright future of this promising line of enquiry is an improved methodology that combines comparable and parallel corpora designed in a larger number of languages, and integrates complementary levels of analysis – textual, cognitive, sociological and cultural – that can be linked together through a causal research model (Chesterman 2005), aiming to cut across interdisciplinary boundaries and go beyond description.

References

Baker, M. (1993) "Corpus linguistics and translation studies. Implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam/Philadelphia, pp. 233-250.

Baker, M. (1996) "Linguistics and cultural studies: complementary or competing paradigms in translation studies?", in A. Lauer, H. Gerzymisch-Arbogast, J. Haller and E. Steiner (eds), *Übersetzungswissenschaft im Umbruch: Festschrift für Wolfram Wilss zum 70. Geburstag*, Gunter Narr, Tübingen, pp. 9-19.

Baker, M. (2001) "Patterns of idiomaticity in translated *vs* original English", paper presented at the Third EST Congress Translation Studies: Claims, Changes and Challenges, August 30 – September 1, 2001, Copenhagen.

Blum-Kulka, S. (1986) "Shifts of cohesion and coherence in translation", in J- House and S. Blum-Kulka (eds), *Inter-lingual and Inter-cultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, Gunter Narr, Tübingen, pp. 17-35.

Blum-Kulka, S., and E.A. Levenston (1983) "Universals of lexical simplification", in C. Faerch and G. Casper (eds), *Strategies in Inter-language Communication*, Longman, London/New York, pp. 119-139.

Chesterman, A. (2000) "A causal model for translation studies", in M. Olohan (ed.), *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*, St. Jerome, Manchester, pp.15-27.

Chesterman, A. (2004) "Beyond the particular", in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 33-50.

Chesterman, A. (2005) "Towards consilience?", in K. Aijmer and C. Alvestad (eds), *New Tendencies in Translation Studies*, Göteborg University, Göteborg, pp. 19-27.

Chesterman, A. and R. Arrojo (2000) "Shared ground in translation studies", *Target* 12 (1), pp. 151-160.

Croft, W. (1990) *Typology and Universals*, Cambridge University Press, Cambridge.

Eskola, S. (2004) "Untypical frequencies in translated language: a corpus-based study on a literary corpus of translated and non-translated Finnish", in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 83-100.

Halverson, S. (2003) "The cognitive basis of translation universals", *Target* 15 (2), pp. 197-241.

Hermans, T. (2006) "Literary translation", in P. Kuhiwczak and K. Littau (eds), *A Companion to Translation Studies*, Multilingual Matters, Clevedon, pp. 77-91.

Kenny, D. (2001) *Lexis and Creativity in Translation. A Corpus-based Study*, St. Jerome, Manchester.

Kenny, D. (2005) "Parallel corpora and translation studies: old questions, new perspectives? Reporting *that* in Gepcolt: a case study", in G. Barnbrook, P. Danielsson and M. Mahlberg (eds), *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, Continuum, London/New York, pp. 154-165.

Kujamäki, P. (2004) "What happens to 'unique items' in learners' translations? 'Theories' and 'concepts' as a challenge for novices' views on 'good translation'", in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 187-204.

Langacker, R. (1987) *Foundations of Cognitive Grammar 1*, Stanford University Press, Stanford, California.

Laviosa, S. (1998) "Core patterns of lexical use in a comparable corpus of English narrative prose", in S. Laviosa (ed.), *L'approche basée sur le corpus/The Corpus-based Approach*, Special Issue of *Meta* 43 (4), Les Presses de L'Université de Montréal, Montréal, pp. 557-570.

Mauranen, A. (2000) "Strange strings in translated language. A study on corpora", in M. Olohan (ed.), *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*, St. Jerome, Manchester, pp. 119-141.

Mauranen, A. (2004) "Corpora, universals and interference", in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 65-82.

Mauranen, A. and P. Kujamäki (eds) (2004) *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia.

Musacchio, M.T. (2005) "The influence of English on Italian: the case of translations of economics articles", in G. Anderman and M Rogers (eds), *In and Out of English: For Better, For Worse?*, Multilingual Matters, Clevedon, pp. 71-96.

Nilsson, P.O. (2004) "Translation-specific lexicogrammar? Characteristic lexical and collocational patterning in Swedish texts translated from English", in A.

Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 129-142.

Olohan, M. (2003) "How frequent are the contractions? A study of contracted forms in the translational English corpus", *Target* 15, pp. 59-89.

Olohan, M. (2004) *Introducing Corpora in Translation Studies*, Routledge, London/ New York.

Olohan, M. and M. Baker (2000) "Reporting *that* in translated English: evidence for subconscious processes of explicitation?", *Across Languages and Cultures* 1 (2), pp. 141-158.

Ǿverås, L. (1998) "In search of the third code: an investigation of norms in literary translation", in S. Laviosa (ed.), *L'approche basée sur le corpus/The Corpus-based Approach*, Special Issue of *Meta* 43 (4), Les Presses de L'Université de Montréal, Montréal, pp. 571-588.

Pápai, V. (2004) "Explicitation: a universal of translated text?", in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 143-164.

Pavesi, M. (2003) "A look 'into' simplification and the translation of motion events in science", in L. Merlini Barbaresi (ed.), *Complexity in Language and Text,* Edizioni Plus – Università di Pisa, Pisa, pp. 147-168.

Robinson, D. (1997) *Translation and Empire: Post-colonial Approaches Explained*, St. Jerome, Manchester.

Rohdenburg, G. (1996) "Cognitive complexity and increased grammatical explicitness in English", *Cognitive Linguistics* 7 (2), pp. 149-182.

Saldanha, G. (2004) "Accounting for the exception to the norm: split infinitives in translated English", in A. Kruger (ed.), *Corpus-based Translation Studies: Research and Applications*, Special Issue of *Language Matters. Studies in the Languages of Africa*, 35 (1), pp. 39-53.

Scarpa, F. (2006) "Corpus-based specialist-translation quality assessment: a study using parallel and comparable corpora in English and Italian", in M. Gotti and S. Šarčević (eds), *Insights into Specialised Translation*, Peter Lang, Bern, pp. 155-172.

Stewart, D. (2000) "Conventionality, creativity, and translated text: the implications of electronic corpora in translation", in M. Olohan (ed.), *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*, St. Jerome, Manchester, pp. 73-91.

Tirkkonen-Condit, S. (2004) "Unique items – over- or under-represented in translated language?", in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 177-186.

Toury, G. (1986) "Monitoring discourse transfer: a test-case for a developmental model of translation", in J. House and S. Blum-Kulka (eds), *Inter-lingual and Inter-cultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies,* Gunter Narr, Tübingen, pp. 79-94.

Toury, G. (1995) *Descriptive Translation Studies and Beyond*, John Benjamins, Amsterdam/Philadelphia.

Toury, G. (2004) “Probabilistic explanations in translation studies: welcome as they are, would they qualify as universals?”, in A. Mauranen and P. Kujamäki (eds), *Translation Universals. Do they Exist?*, John Benjamins, Amsterdam/Philadelphia, pp. 15-32.

Tymoczko, M. (1998) “Computerized corpora and the future of translation studies”, in S. Laviosa (ed.), *L’approche basée sur le corpus/The Corpus-based Approach*, Special Issue *Meta* 43 (4), pp. 652-659.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 239-255 © 2008 Polimetrica International Scientific Publisher Monza/Italy

"What's in a name?" References to women in *Romeo and Juliet* and their translation into Italian

Vincenza Minutella – University of Turin

1. Introduction

This paper aims to identify and describe the words used by Shakespeare's characters in *Romeo and Juliet* to refer to women, with particular reference to the word *MAID*, and to analyse translation strategies for this word in different versions for the page and the stage. The various problematic issues which translating a theatre text involves are the main background against which this analysis is carried out.

In the famous rhetorical question uttered by Juliet during the balcony scene (Act II, scene 2), "What's in a name?", Shakespeare seems to anticipate the question raised by twentieth-century linguists about lexical meaning. A basic assumption of modern linguistics (Halliday 1989) is that the co-text and context in which any text or utterance is inserted are of fundamental importance in understanding its meaning. Words do not live in isolation but "enter into meaningful relations with other words around them" (Sinclair 1996: 76), which may add a particular 'colouring' to the word in question (Louw 1993). The co-text influences the use and connotation of a word, and it is therefore useful to examine the combinatory patterns of a word in order to understand its particular meaning(s). However, the focus on short stretches of language may not always explain the full range of meanings that a word may carry, so that longer stretches, full texts, and contextual information, as well as subjective observations may sometimes be needed to provide a comprehensive analysis.

Descriptive Translation Studies argue for empirical research and for the need to insert the translations in the wider socio-cultural context in which they are produced and have to function. In this way specific translation choices can be understood in the light of extra-linguistic and cultural factors that affect the translation (see Bassnett 1998). Translation choices may therefore be affected by the specific historical period in which the translator works, by the medium – i.e. whether the target text is to be published, or to be performed on stage – by the function of the translation and the interpretation of the director, and by the audience or readers it is aimed at.

As far as the difficulties of translating theatre texts are concerned, *Romeo and Juliet* constitutes a perfect example, since the translation of this play poses several problematic issues. First, as it is an Elizabethan text, the translator needs to choose between a historicising/foreignising or a modernising/domesticating approach to language. The former implies taking the readers towards the source text, while the latter takes the source text towards the target readers, by using modern words and expressions, and simplifying aspects which have become obsolete or difficult to understand (Schleiermacher [1813], 1992; Venuti 1998). The fact that it is a theatre text written to be performed constitutes a second difficulty: the words on the page should be 'playable', and theatre texts are very often cut in production. A translation made for publication and one made for the stage may apply different translation strategies, and the theatre tends to accept a higher degree of cutting and 'adaptation' of the source text, due to directorial view and the need to make the text communicate to a contemporary audience (see for instance Heylen 1993; Aaltonen 2000).

Romeo and Juliet lends itself to the exploration of the ways women are addressed and referred to in different speeches and by different characters, as the play centres around the theme of love, and portrays different relationships between men and women. The presence of complex wordplay and vulgar allusions in *Romeo and Juliet* poses a further difficulty for translators, as the taste of the time or the target readers/audience may not accept references to sex.

The analysis of three different translations of the play by different authors in different historical periods will show how this problem has been dealt with. The aim of the comparison between the English source text and its Italian target texts is to ascertain the influence of the medium and the function of the translation (whether for the page or a stage performance) and of the time factor (whether a nineteenth or twenty-first-century version) on translation strategies. The analysis will also attempt to identify patterns in translation choices.

2. Reference texts and methodology

As regards the reference texts, the choice of *Romeo and Juliet* raises problems as there are several editions of the play, some of which can be downloaded from the Internet.[1] The electronic version available on the website Open Source Shakespeare,[2] which contains Shakespeare's Complete Works (from the 1864 Globe Edition) was downloaded and saved in .txt format.

As for the translations, in order to have a variety of target texts in terms of time and scope, Pietro Deandrea and Marco Ponti's 2005 translation for the stage,[3] and two published versions have

[1] The electronic version of the First Folio edition, from the Oxford Text Archive, was not chosen because its spelling is different from modern editions, and could create problems in searching for words. Another option was to download an edition with modern spelling from the many available on the web, such as the Oxford Shakespeare edition, edited by W. Craig, 1914 (http://bartleby.com/70/), the Project Gutenberg version (http://www.gutenberg.org), the one available on the website The Complete Works of William Shakespeare (http://shakespeare.mit.edu), or the version available on the website Shakespeare's words (http://www.shakespeareswords.com), from the Penguin edition of the plays. Most of the versions needed mark-up before being inserted in the programme WordSmith Tools 4.0, which could render the analysis of concordances more time-consuming (see Culpeper 2002: 14-15 for a discussion of problems related to the choice and preparation of the text).

[2] The website includes a concordance programme which enables users to search for specific words. Concordances contain the co-text as well as information on who uses the word, and in what acts and scenes.

[3] Deandrea and Ponti's translation was commissioned by theatre director Gabriele Vacis. The translation is contained in the volume *William Shakespeare. Romeo&Juliet. Un progetto di Gabriele Vacis e Roberto Tarasco. Regia di Gabriele*

been selected: the printed translation by Carlo Rusconi (1852) and Salvatore Quasimodo's translation, edited by Giorgio Melchiori (1976, first edition in 1949). While Rusconi's version is not very well-known, Quasimodo's translation has been published, reprinted and re-edited by Mondadori for several years, and is quite popular.

As far as the methodology is concerned, the retrieval and analysis of data has been conducted with the aid of the WordSmith Tools 4.0 software as well as the concordancer available on http://www.opensourceshakespeare.org, which provides longer stretches of co-text and information about acts, scenes and speakers. The main words used to talk about women have been identified in terms of frequency, and concordances for such words help to highlight the patterns in which these words occur, and the meanings they acquire in the play. For the translations the analysis has been carried out manually.

3. Analysis of Data

Romeo and Juliet contains several words referring to women: *WOMAN* and its synonyms or near synonyms *GIRL, LADY*, *MAID, MAIDEN, MADAM, MISTRESS,* in their singular and plural forms, and in compounds. The following table illustrates the most frequent 'women' words in *Romeo and Juliet*. It is evident that *MAID* is the most morphologically productive and the most amenable to ambiguity and to wordplay. As it appears to be particularly interesting in terms of combinatory patterns and/or variations in meaning, *MAID* in its singular, plural and combined forms will be discussed in more detail in the following section.

Vacis, Fondazione del Teatro Stabile di Torino, Torino, published in 2005. I am grateful to Pietro Deandrea for allowing me to use his translation.

'Women' words in Romeo and Juliet	**Frequency**
LADY	52
LADIES	5
MADAM	23
MISTRESS	11
MISTRESSES	1
MAID	10
MAIDS	6
CHAMBER MAIDS	1
MAIDEN	1
MAIDENHEAD	2
MAIDENHEADS	1
MAIDENHOODS	1
MAIDEN-WIDOWED	1
GIRL	6
GIRLS	1
WOMAN	3
WOMEN	4
GENTLEWOMAN	4

Table 1. Frequency of words referring to women in *Romeo and Juliet*.

MAID

According to the *Oxford English Dictionary* the word *MAID* originated in Middle English and has 4 main meanings as a noun referring to women: 1. a girl or a young unmarried woman (now archaic, poetic, or playful), 2. a virgin, 3. an unmarried woman, a spinster (now rare), 4. a female servant or attendant, often with defining word prefixes.

In *Romeo and Juliet*, the word *MAID* is used with various meanings, and in combination with other words: it occurs 10 times in the singular, 6 in the plural, and in lexical items such as *MAIDEN* (1), *MAIDENHEAD(S)* (3), *MAIDENHOODS* (1), *MAIDEN-WIDOWED* (1) and *CHAMBER MAIDS* (1). In total, there are 23 occurrences of the word, which are shown in the concordance lines below (to which the name of the speaker has been added in brackets):[4]

1 night is on my face, Else would a **maiden** blush bepaint my cheek (J)
2 you for a highway to my bed; But I, a **maid**, die maiden-widowed. (J)
3 little worm Prick'd from the lazy finger of a **maid**; Her chariot (M)
4 mother much upon these years That you are now a **maid.** Thus then (LC)
5 his mistress were that kind of fruit As **maids** call medlars, when(M)
6 Cophetua loved the **beggar-maid**! He heareth not, he stirreth not(M)
7 here will I remain With worms that are thy **chamber-maids**; O, here (R)
8 bed; But I, a maid, die **maiden-widowed.** Come, cords, come, Nurse(J)
9 Heaven and yourself Had part in this fair **maid;** now heaven hath(FL)
10 pale with grief, That thou her **maid** art far more fair than she: (R)
11 far more fair than she: Be not her **maid**, since she is envious (R)
12 Montague's men from the wall, and thrust his **maids** to the wall. (S)
13 Now, by my **maidenhead**, at twelve year old, I bade her come. (N)
14 I'll to my wedding-bed; And death, not Romeo, take my **maidenhead (J)**
15 to stand: I will take the wall of any man or **maid** of Montague's. (S)
16 Your lady's love against some other **maid** That I will show you (B)
17 lose a winning match, Play'd for a pair of stainless **maidenhoods** (J)
18 men, I will be cruel with **the maids**, and cut off their heads. (S)
19 Ay, the heads of **the maids**, or their maidenhenheads; Take it (S)
20 Gregory: The heads of **the maids**? (G)
21 And all the better is it for **the maid**: Your part in her you (FL)
22 heads of the maids, or their **maidenheads**; take it in what sense (S)
23 This is the hag, when **maids** lie on their backs, That presses them(M)

The analysis of the concordances above, and of the wider co-text and context – the speaker, and the situation in which the word is uttered – highlights some grammatical and lexical patterns. *MAIDEN* appears only once in the play, it is uttered by Juliet in the 'balcony scene' (II, 2), and it is used as an adjective ("a maiden blush") meaning 'unmarried, virgin, pertaining to a young unmarried woman'. In the play *MAID* has two main meanings: it is used to refer to a young unmarried woman or a virgin (examples 2 to 5, 9, 16, 21 and 23), and as a synonym of female servant or attendant (examples

[4] J stands for Juliet, M for Mercutio, R for Romeo, LC for Lady Capulet, C for Capulet, B for Benvolio, S for Sampson, G for Gregory, N for the Nurse, FL for Friar Laurence.

6, 7, 10-12, 15, 18, 19, 20). However, as will be shown, in examples 12, 15, 18, 19, 20 the referent may be ambiguous, it could be both servants and young women. In terms of grammatical preference, *MAID* is often preceded by determiners, in particular by the possessive adjective in its sense of servant. As highlighted by the dispersion plot, the word (in its singular, plural and in a compound) is particularly frequent in the opening scene of the play, in the exchange between Capulet’s servants Gregory and Sampson (I, 1). Here, *MAID* is inserted in a discourse where negative references to sex, violence and vulgar puns prevail:

> SAMPSON: A dog of that house shall move me to stand: I will take the wall of any man or **maid** of Montague's.
>
> GREGORY: That shows thee a weak slave, for the weakest goes to the wall.
>
> SAMPSON: ‘Tis true; and therefore women, being the weaker vessels, are ever thrust to the wall: therefore I will push Montague's men from the wall, and thrust his **maids** to the wall.
>
> GREGORY: The quarrel is between our masters and us their men.
>
> SAMPSON: 'Tis all one, I will show myself a tyrant: when I have fought with the men, I will be cruel with the **maids**, and cut off their heads.
>
> GREGORY: The heads of the **maids**?
>
> SAMPSON: Ay, the heads of the **maids**, or their **maidenheads**; take it in what sense thou wilt.

The meaning of *MAID(S)* in the above occurrences is ambiguous, as the men may talk both of female servants and of young women. The above exchange also contains the compound *MAIDENHEAD*, which is an archaic word for virginity, or the condition of a maiden, as well as the hymen (OED). Sampson’s wordplay (“heads of the maids” – “maidenheads”) alludes to virginity and the hymen, and the vulgar pun is quite explicit. The servant’s bawdy puns, which constitute the first lines of the play, contain sexual allusions and hints at violence towards women, which may aim at making explicit

> at the beginning of this love tragedy, one possible relationship between man and woman: a brutal male dominance expressed in sadistic quibble. (Mahood 1979: 60).

Another character who uses the word *MAID* in a negative co-text, accompanied by references to sex and violence, is Mercutio. In the 'Queen Mab speech' (I, 4), he says

> this is the hag, when **maids** lie on their backs, / That presses them and learns them first to bear, / Making them **women** of good carriage.

In their nightmares maids learn to behave as sexual objects that are taken by force by men, and learn to become bearers of children, real women. His use of *MAIDS* in Act II, scene 1, is another example of his attitude towards women and exploitation of sexual innuendo:

> Now will he sit under a medlar tree, / And wish his **mistress** were that kind of fruit / As **maids** call medlars, when they laugh alone. / Romeo, that she were, O, that she were / An open et caetera, thou a poperin pear!

Mercutio associates *MISTRESS* and *MAIDS* with the word *MEDLARS* (a fruit which was thought to resemble the female genitalia), and adds a vulgar quibble on *MEDLAR/MEDDLER*, as *TO MEDDLE* means to have sexual intercourse. The following *POPERIN PEAR* also hints at sexual organs. The whole speech in II, 1 contains several sexual allusions and puns. All Mercutio's references to women in the play (through the words *MAID*, *WOMEN* and *MISTRESS*) betray a negative attitude towards them. In fact, these words seem to be always associated with others that carry negative connotations or that belong to the domain of sex and violence, and are often accompanied by vulgar puns. Sometimes his attitude is also ironic, as in his use of puns in the above quotation.

As regards the compounds, the other two occurrences of *MAIDENHEAD* are uttered by the Nurse ("Now, by my maidenhead, at twelve year old") and by Juliet. This shows how sexual allusions are also made by women in the play. The Nurse is usually considered quite vulgar, while Juliet is actually very much aware of her sexuality. The girl even seems to play on the word *MAID* when she hears about Romeo's banishment (III, 2):

> Take up those cords: poor ropes, you are beguiled, / Both you and I; for Romeo is exiled: / He made you for a highway to my bed; / But I, a **maid**, die **maiden-widowed**. / Come, cords, come nurse; I'll to my wedding-bed; / And death, not Romeo, take my **maidenhead**!

Juliet is also the only one in the play who uses the word *MAIDENHOOD* - which means the condition of being a virgin. In the 'Gallop apace' monologue (III, 2), when she is waiting for Romeo in her room, she says: "Come, civil night, / Thou sober-suited matron, all in black, / And learn me how to lose a winning match, / Play'd for a pair of stainless **maidenhoods**."

4. The translation of *MAID*

Manual analysis of the translation of the word *MAID* by the three translators for page and stage has revealed interesting patterns and the tendency to adopt three main translation strategies: full translation, omission and adaptation/rewriting.

Rusconi's translation (1852)

The most striking aspect of Rusconi's translation of *MAID(S)* is the presence of several omissions (9 out of the 23 occurrences are omitted). These are more frequent in Mercutio's speeches, as Rusconi cut many lines from them. For instance, in the famous 'queen Mab speech' (I, 4), Mercutio's "Her wagoner a small grey-coated gnat, / Not so big as a round little worm / Prick'd from the lazy finger of a **maid**" as well as the last lines of the monologue, "This is the hag, when **maids** lie on their backs, / That presses them and learns them first to bear, / Making them women of good carriage" are not translated at all. Such excisions are not justified by the translator. However, in a footnote at the beginning of the 'Queen Mab speech' he explains that: "Per quanto strana e inopportuna possa parere questa descrizione dei sogni ai lettori italiani, essa gode in Inghilterra della più alta celebrità" (1852: 169). The translator thus omitted lines from this monologue because he considered them unsuitable to the taste of his Italian audience. Rusconi's translation of Mercutio's vulgar speech in II, 1 also tends to adapt and tone down vulgar allusions, and the lines containing the word *MAID* are cut:

> Now will he sit under a medlar tree, / And wish his **mistress** were that kind of fruit / As **maids** call medlars, when they laugh alone. / Romeo, that she were, O, that she were / An open et caetera, thou a poperin pear!
>
> Ah! Senza dubbio egli ora se ne starà assiso sotto qualche antico salice, per esalarvi fra l'aure gl'insensati suoi voti, e porger preci affinché la sua **fanciulla** si renda flessibile come i rami che gli stan sopra (1852: 174-175).

The last lines in particular highlight a certain degree of censorship and rewriting: "and wish his mistress were that kind of fruit as **maids** call medlars" is omitted and adapted, and vulgar allusions to male and female organs are eliminated. It should also be noted that the translator felt the need to add a footnote at the beginning of the scene, in which he explained that he did not translate some lines, "che non istimammo conveniente di tradurre" (1852: 174).

In I, 1 Sampson's pun "the heads of the **maids**, or their **maidenheads**. Take it in what sense thou wilt" becomes "Sì, la **lor** testa, se riscattarla non vorranno col dono che glie ne chiederei", in which *MAIDS* is translated through a pronoun, and there is no mention of virginity. In this opening scene the repeated words *MAID/S* are also translated with *FANCIULLE* (3 occurrences) and *FEMMINE* (once), therefore opting for the meaning of 'young women', not of 'servants'. The choice of *FEMMINE* is interesting as the word makes reference to sexuality, and as a synonym for woman it tends to be derogatory. Rusconi's translation "mi accontenterò di stendervi sopra le loro belle **femmine**" renders Sampson's view of women as sexual objects that are taken by force ("I will thrust his **maids** to the wall"). This is perhaps the only case in which Rusconi does not censor the text.

His omission of lines containing sexual allusions and bawdy language, which is probably due to his moral views, is evident in his translation of *MAIDENHEAD(s)* and *MAIDENHOODS*, which are always eliminated. For instance, the Nurse's "by my **maidenhead**, at twelve year old" (I, 3) is rendered with the more neutral "sull'onor mio". Here and in a few other cases Rusconi adds a footnote that explains what he had cut.[5] Juliet's words "But I, a **maid**, die **maiden-widowed**. / Come, cords, come, nurse; I'll to my wedding-bed; / And death, not Romeo, take my **maidenhead**" (III, 2) are rendered with "Muori dunque, **vedova vergine**. Andiamo, nutrice: vo' coricarmi sul mio letto nuziale, che in breve sarà fatto mia bara", with a footnote quoting the original sentence and its literal translation.

Other interesting translations of *MAID/S* are the paraphrases *ALL'ETÀ VOSTRA*, which renders Lady Capulet's "I was your mother much upon these years / That you are now a **maid**" ("ed io mi ricordo che ero già madre all'età vostra") and the formal "vergine del suo culto", uttered by Romeo in the balcony scene: "That thou, her **maid**, art far more fair than she" ("che tu, **vergine del suo culto**, splenda più chiara di lei"). Rusconi also uses pronouns to replace the noun, which is a kind of rewriting that permits repetition to be avoided. For instance, Friar Laurence's "Heaven and yourself / Had part in this fair **maid**; now heaven hath all, / And all the better is it for the **maid**" (IV, 5) becomes "il Cielo e voi avevan parte di quella **fanciulla**, che ora il Cielo solo possiede; ed è ventura per **lei**." The Italian *FANCIULLA/E* occurs 4 times to render the word *MAID/S*.

Quasimodo's translation (1976)

On the contrary, omission in Quasimodo only occurs when the meaning can be understood from the co-text, and to avoid repetition. For this purpose, also pronouns and paraphrase tend to be used. For instance, Romeo's "that thou, her **maid**, art far more fair than she. Be not her **maid**" becomes "che ha invidia di te perché sei bella più di lei, **tu che la servi**. E se ha invidia di te lasciala sola." Friar Laurence's words, containing a repetition of *MAID*, are rendered as follows: "questa bella **fanciulla** era vostra e del cielo; ora è tutta del cielo, ed è la cosa più bella **per lei**." However, the most relevant aspect of Quasimodo's translation is his tendency to retain the full text, also accepting repetitions. The different occurrences of *MAID* are expressed by the words *SERVA/SERVE* (2 occurrences in the first

[5] The footnote explains: "il testo porta: by my maidenead, at twelve year old; cioè a dire per la mia verginità a 12 anni" (1852: 165).

scene), *FANCIULLA/E* (4 occurrences), *RAGAZZE* (5 occurrences) and with pronouns and paraphrase. As shown in the translation of Gregory and Sampson's dialogue (I, 1), wordplay and sexual allusions are rendered, and all the occurrences of *MAID/S* are translated using two meanings:

> SANSONE: Dico che un cane di quella casa mi ecciterà a star fermo. Avrò il lato del muro da qualunque servo, ed anche **serva**, di casa Montecchi che incontrerò. […] SANSONE: Verissimo; e per questo le donne, che sono i vasi più deboli, sono spinte sempre contro il muro. Caccerò, dunque, via dal muro i servi del Montecchi e forzerò al muro le sue **serve**. […] quando mi sarò battuto con gli uomini, sarò duro con **le ragazze** e le sferzerò tutte. GREGORIO: Sferzare **le ragazze**? SANSONE: Sì, **sferzare o sforzare le ragazze**. Prendilo nel senso che vuoi.

As shown by the following examples, Mercutio's comments about women are also kept: "grande meno della metà del verme che gonfia il dito alle **fanciulle** pigre" ("not so big as a round little worm prick'd from the lazy finger of a **maid**", I, 4); "Mab è la strega che se trova supine **le ragazze** le costringe all'abbraccio" ("This is the hag, when **maids** lie on their backs, that presses them and learns them first to bear", I, 4); "sogna con desiderio la sua donna; la vede nella forma di quel frutto che **le ragazze** ridendo chiamano 'nespola' quando son sole" ("and wish his mistress were that kind of fruit as **maids** call medlars when they laugh alone"). Juliet and the Nurse's references to *MAIDENHEAD* and *MAIDENHOOD* are also fully translated: "Ve lo giuro sulla **verginità** di quando avevo dodici anni" (I, 3, Nurse); "dove si giocano due **verginità** intatte", "Ma io, ancora **fanciulla**, morirò **vergine e vedova**. […] la morte, non Romeo, prenderà la mia **verginità**" (III, 2, Juliet).

Deandrea and Ponti's translation (2005)

The most striking aspect of Deandrea and Ponti's translation for the stage seems to be a high degree of adaptation which renders vulgar allusions more explicit. The dialogues are often rewritten in a more modern, colloquial and direct style, some words are eliminated but some are also added, and puns are adapted to contemporary language. Some lines from Mercutio's speech (II, 1) illustrate this comment:

> This cannot anger him. 'Twould anger him / To raise a spirit in his **mistress**' circle / Of some strange nature, letting it there stand / Till she had laid it and conjured it down: / That were some spite. My invocation / Is fair and honest; in his **mistress**' name / I conjure only but to raise up him. […] Now will he sit under a medlar tree / And wish his **mistress** were that kind of fruit / As **maids** call medlars when they laugh alone. / O Romeo, that she were, O that she were / an open et caetera and thou a poperin pear!
>
> Mica ho evocato uno di quegli spiriti dispettosi che entrano nelle mutande delle **ragazze** e se ne stanno lì dentro fino a che non si ammosciano. No, il mio incantesimo è buono e puro. Uso il nome della sua **amata** così gli viene…voglia, e torna qui. […] già me lo vedo, sdraiato sotto un albero di fico, a pensare a quella parte di **Rosaline** che starebbe benissimo in mezzo a quei frutti. Oh, Romeo! Se solo lei si decidesse a dartela, la sua virtù.

In the above example there is the omission of 'as maids call medlars', but there is adaptation whose result is that the words are clearly vulgar and sexual puns are rendered more explicit and colloquial (the change from "medlar tree" to "albero di fico", the use of "si decidesse a dartela, la sua virtù"). Notice also the vulgar puns associated with the word MISTRESS. Mercutio's last lines of the 'queen Mab speech' (I, 4) also contain adaptation: "lei che alle volte corre lungo i ventri nudi delle **ragazze** la notte, e insegna loro quello che tutte le **ragazze** devono sapere" ("this is the hag, when **maids** lie on their backs, that presses them and learns them first to bear, making them **women** of good carriage"). Adaptation is also present in the rendering of the first scene, where *MAID/S* are translated with *DONNE* as well as with pronouns which are used to avoid repetition:

> SAMPSON: Comunque, se arriva un qualunque cane Montecchi, io sono pronto. Uomini o **donne**, sai cosa? Per me nessuna differenza. GREGORY: Questa è una scelta di vita tua personale, io non ci voglio entrare. SAMPSON: Che dici? Per gli uomini, la spada, ma per **le donne** ho in mente altro. GREGORY: Lascia stare le loro **donne**, è una lite tra uomini, questa. Padroni e servi, ma solo uomini. SAMPSON: Mica **le** voglio uccidere, per chi mi prendi. Me **le** voglio solo fare, tutte quante, senza pietà, **le** infilzo contro il muro.

In the above example the pun *HEADS OF THE MAIDS-MAIDENHEADS* is omitted, but adaptation and explicitation render the sexual allusion very clear, and add a hint at homosexuality.

Deandrea and Ponti also resort to omission, and this may be due to the fact that they eliminate some lines from the source text, and render exchanges more direct and shorter. For instance, Lady Capulet's "that you are now a **maid**" is omitted, as her dialogue with Juliet about Paris is shortened and reduced to a modern-day exchange: "LADY CAPULET: Dimmi, che ne pensi dell'idea del matrimonio? JULIET: Neanche per sogno, mamma. LADY CAPULET: Il nobile Paris è innamorato di te."[6] Juliet's words in III, 2 also provide an example of translation by omission, as they are shortened: "Prendi quelle corde, per favore. Poverine, che beffa anche per loro. Dovevano portarlo fino al mio letto, e invece mi vedranno morire come **una vedova vergine**." Despite cuts and a modernising approach, this translation retains the vulgar references to women contained in the play.

5. Conclusion

The comparison of different translations of *Romeo and Juliet* has identified some repeated or common patterns in the translation of words referring to women: a tendency to avoid repetition, translation of the full text, omission and adaptation. Some translators prefer one strategy over others. Possible reasons for such tendencies are to be found in the context of production of the target texts.

As illustrated by the examples above, translators in most of the cases – whether they translate for the page or the stage – tend to avoid repetition and adopt paraphrase, synonyms, pronouns or words belonging to the same semantic field. However, repetitions seem to be tolerated more in translation for the page. Quasimodo's version, which was published in the mid-twentieth century, translates Shakespeare's text in its entirety, as it avoids omissions.

[6] Compare with the original: "LC: Tell me, daughter Juliet, / How stands your disposition to be married? J: It is an honour that I dream not of. […] Well, think of marriage now. Younger than you, / here in Verona, ladies of esteem, / Are made already mothers. By my count, / I was your mother much upon these years / that you are now a **maid**. Thus then in brief: / the valiant Paris seeks you for his love." (I, 3)

The strategies of omission and adaptation are instead largely adopted by Deandrea and Ponti – who translated for the stage in recent years – and by Rusconi – who translated for the page in the nineteenth century. In Rusconi's version significant alterations are made to the source text through translation which cuts or alters vulgar allusions and wordplay. Sometimes the translator explains the reasons for his omissions, whereas in several cases the lines are simply excised or changed. Rusconi's footnotes reveal that his personal poetics and the taste of his time dictated his translation choices, so that the parts that were perceived as immoral or exaggerated were eliminated, or adapted. His omissions and rewriting of the source text can be considered examples of censorship.

Conversely, in the case of Deandrea and Ponti's contemporary translation for the stage, the reasons for omissions are more practical and they are related to the medium and to a specific directorial view. Contemporary theatre conventions require that performances last approximately two hours, while the full text of the play would take much longer to be performed. The director's interpretation of the play also influences translation strategies and text cutting. Deandrea and Ponti apply the strategy of adaptation, which enables the translators to express a concept more directly, to explain or simplify it and to adapt the language to the target audience. They adopt a modernising approach, as the language used in the target text is contemporary and colloquial. The translation contains vulgar explicit language, and his rendering of words referring to women is openly sexual. This is in line with the director's interpretation of the characters in the play as contemporary 'vitelloni trentenni' that speak a colloquial language. As Ponti and Deandrea explain, discussing their approach to translation and Vacis' view of the play, "il testo viene leggermente accorciato, non tanto per ragioni di tempo scenico quanto per privilegiare gli aspetti della vita un po' da 'vitelloni' di Romeo, Mercuzio e compagnia" (2005: 25). As a result, comedy and vulgar puns were highlighted in this translation and in performance.

By comparing the translation of the word *MAID* in different target texts, this paper has suggested that the reasons for the tendencies discussed above (avoidance of repetition, full translation, omission and adaptation) are to be searched for in the target context in which the translations are produced, in the medium of transmission and the function of the target text in the receiving culture. By describing one source text and a small number of its translations, and by focussing only on one specific word and its semantic field, this analysis has been inevitably limited. However, it points to some possible areas of research and illustrates how the extraction of corpus data and the study of lexis can be usefully combined with translation studies.

References

Aaltonen, S. (2000) *Time-Sharing on Stage. Drama Translation in Theatre and Society*, Multilingual Matters, Clevedon.

Baker, M. (1993) “Corpus linguistics and translation studies: implications and applications”, in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam/Philadelphia, pp. 233-250.

Baker, M. (ed.) (1998) *Routledge Encyclopedia of Translation Studies*, Routledge, London/New York.

Bassnett, S. (1998) “The translation turn in cultural studies”, in S. Bassnett and A. Lefevere, *Constructing Cultures. Essays on Literary Translation*, Multilingual Matters, Clevedon, pp. 123-140.

Blakemore Evans, G. (ed.) [1984] (1998) *Romeo and Juliet*, The New Cambridge Shakespeare, Cambridge University Press, Cambridge.

Culpeper, J. (2002) “Computers, language and characterisation: an analysis of six characters in Romeo and Juliet”, in U. Melander Marttala, C. Ostman and M. Kyto (eds), *Conversation in Life and in Literature: Papers from the ASLA Symposium, Association Suedoise de Linguistique Appliquée (ASLA)*, 15, Universitetstryckeriet, Uppsala, pp. 11-30. http://www.lexically.net/wordsmith/corpus_linguistics_links/Keywords-Culpeper.pdf

De Mauro, T. (2007) *Dizionario della lingua italiana De Mauro-Paravia* (on-line version).

Deandrea, P. and M. Ponti (2005) “Romeo & Juliet di William Shakespeare. Raccontato da Marco Ponti e Pietro Deandrea”, in *William Shakespeare. Romeo & Juliet, R&J Links, un progetto di G. Vacis e R. Tarasco*, Fondazione del Teatro Stabile di Torino, Torino, pp. 87-157.

Devoto, G. and G.C. Oli (2007) *Vocabolario della lingua italiana Devoto Oli*, Le Monnier, Firenze.

Gibbons, B. (ed.) [1980] (1998) *Romeo and Juliet*, The Arden Shakespeare, Routledge, London.

Halliday, M. (1989) *Language, Context and Text: Aspects of Language in a Social Semiotic Perspective*, John Benjamins, Amsterdam.

Heylen, R. (1993) *Translation, Poetics, and the Stage. Six French Hamlets*, Routledge, London/New York.

Hoenselaars, T. (ed.) (2004) *Shakespeare and the Language of Translation*, The Arden Shakespeare, London.

Laviosa, S. (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*, Rodopi, Amsterdam/New York.

Levenson, J. (ed.) (2000) *Romeo and Juliet,* The Oxford Shakespeare, Oxford University Press, Oxford.

Louw, B. (1993) "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies", in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam/Philadelphia, pp. 157-176.

Macmillan English Dictionary for Advanced Learners (2002), Macmillan, Oxford.

Mahood, M.M. (1979) "Romeo and Juliet", in *Shakespeare's Wordplay*, Methuen, London/New York, pp. 56-72.

Picchi, F. (1999) *Grande dizionario inglese-italiano, italiano-inglese*, Hoepli, Milano.

Ponti, M. and P. Deandrea (2005) "La terribile attualità di Shakespeare", in *William Shakespeare. Romeo & Juliet, R&J Links, un progetto di G. Vacis e R. Tarasco*, Fondazione del Teatro Stabile di Torino, Torino, pp. 25-28.

Quasimodo, S. (1976) *Romeo e Giulietta,* in G. Melchiori (ed.), *Teatro completo di William Shakespeare.* Vol. IV: *Le tragedie*, Mondadori, Milano.

Rusconi, C. (1852) *Teatro completo di Shakespeare. Voltato in prosa italiana da Carlo Rusconi*, Cugini Pomba e comp. Editori, Torino.

Schleiermacher, F. [1813] (1992) "On the different methods of translating", in R. Schulte and J. Biguenet (eds), *Theories of Translation. An Anthology of Essays from Dryden to Derrida*, University of Chicago Press, Chicago/London, pp. 36-54.

Scott, M. (2005) "The behaviour of keywords (KWs)", paper presented at the Corpus Linguistics 2005 Conference.
www.lexically.net/downloads/writing/bham/talk.ppt

Simpson, J.A. and E.S.C. Weiner (1989) *The Oxford English Dictionary on Historical Principles*, second edition, Clarendon Press, Oxford.

Sinclair, J. (1996) "The search for units of meaning", *Textus* IX, pp.75-106.

Toury, G. (1995) *Descriptive Translation Studies and Beyond*, John Benjamins, Amsterdam/Philadelphia.

Venuti, L. (1998) "Strategies of translation", in M. Baker (ed.) *Routledge Encyclopedia of Translation Studies*, Routledge, London/New York, pp. 240-244.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 257-273 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Towards a corpus-based distinction between language-specific and universal features of mediated discourse

Margherita Ulrych – Catholic University of Milan
and Simona Anselmi – Catholic University of Piacenza*

1. Introduction: forms of mediated discourse

Translation is generally recognized as involving some form of mediation be it in an oral or written medium "between two parties for whom mutual communication might otherwise be problematic" (Hatim and Mason 1990: 223). There is, however, no definite agreement as yet as to what the term 'translation' actually covers in terms of communication. Although the notion that translation is simply "a process of linguistic recoding that should aim at a maximal transfer of source text syntax and semantics into the target language" (Delabastita 1989: 213) has largely been surmounted, there is still some resistance in accepting Steiner's (1975: 47) view that "inside or between languages human communication equals translation". There are, of course, a host of other definitions in between and the overriding principle that characterizes all of them is that the concept of translation encompasses factors and processes that were once considered as lying beyond what Jakobson termed "translation proper" (1966: 233). Central in this regard is the translator's mediating presence (Ulrych 1999). Although translation is, generally-speaking, regarded as a social communicative process that takes place within well-defined

* The paper reflects the collaborative work of both the authors: Margherita Ulrych is responsible for sections 1, 2 and 6 and Simona Anselmi for sections 3, 4 and 5.

cultural contexts and specific spatial and temporal settings, Chesterman (1993: 9) is not alone in viewing the translation process in relative terms largely “determined by the translator, on the basis of his or her understanding of the intentions of the original writer and/or commissioner, the type and skopos of the text, and the nature of the prospective readership”. More recently, Pym (2007: 165) affirms that formal conceptualizations of translation “allow considerable space for the historical study of textuality, receptive positions, and the semi-concealed subjectivities of translators”.

The variety of mediating activities that translators are required to undertake in the course of their profession has been extensively described by Sager (1994: 178), ranging from producing a reduced translation from a full source text, such as bilingual abstracting, to creating a full target text on the basis of a minimal draft source text. A distinction is thus made between ‘autonomous’ and ‘dependent’ target documents on the basis of their relationship with the source text. In translation studies mediation is indirectly investigated by Lefevere (1992) through the idea of ‘rewriting’ in the domain of literature. Rewriting for Lefevere is a broad phenomenon that spans all forms of text production comprising “translation, historiography, anthologizing, criticism and editing” (1992: 9). According to Lefevere the general public or what he calls “non-professional readers” do not read literature as written by its writers but as rewritten by its rewriters. These various forms of rewriting are therefore to be interpreted as socialization processes by means of which texts are made more accessible to given audiences. If the concept of rewriting is applied to a wider variety of discourse genres, it may be seen to exist intralingually, interlingually and intersemiotically, to use Jakobson’s classic tripartite definition (1966: 233). Intralingually, rewriting covers not only the compiling of anthologies, abridged or simplified versions of novels and editing but also the popularization of research articles for the general public, which are rewritten in such a way as to match the target audience’s assumed level of comprehension. Interlingually, translations make texts accessible to speakers of other languages and members of other cultures. In today’s globalized world a vast amount of knowledge and experience is mediated through various modes of translation that encompass a host of domains from the scientific, economic and political to philosophy and literature, thus affecting almost every area of human communication. Intersemiotically, rewritings include the transfer of meaning by means of more than one semiotic system, and ranges from theatre and film translation to media translation and from translating cartoons and comic strips to translating an illustrated user's manual. For a considerable majority of people rewritings function as 'originals', for the most part, if not all, of their lives. It is not surprising, therefore, that the issue of how information and events are filtered and relayed, how, in a word, they are 'mediated' through rewriting is a matter for serious investigation.

The present paper starts from the premise that what is considered 'original' communication may in fact be envisaged as mediated, including "language used by non-native speakers, native speakers addressing non-native speakers and editorial intervention" (Ulrych 1999: 38). The whole issue of what constitutes an original text may be called into question as Paz convincingly argues:

> No text can be completely original because language itself, in its very essence, is already a translation - first from the nonverbal world, and then, because each sign and each phrase is a translation from another sign, another phrase. However, the inverse of this reasoning is also entirely valid. All texts are originals because each translation has its own distinctive character. Up to a point, each translation is a creation and thus constitutes a unique text. (Paz 1992: 154)

In particular, it investigates written English as a Lingua Franca (ELF) used by native and non-native speakers within professional and institutional settings as instances of mediated communication and compares them to translational language. The aim is to verify whether the various forms of mediated discourse have features in common and, if so, whether these shared features are universal across languages or whether they are the result of transfer or interference between language-specific pairs. As we shall see more fully below, Jenkins (2006) has convincingly argued in favour of cross-language factors in the area of ELF while, in the area of corpus-based translation studies, Baker (1993) has posited the existence of universal characteristics of translated text whatever the source language. A further, and at the same time central, component of mediated discourse in our study is represented by a corpus of draft texts written in English by non-native speakers and subsequently edited by native speakers.

The initial results of our investigation into written English mediated discourse may also be fruitfully compared to the findings of computer learner corpus research in order to verify whether lingua franca varieties of English are emerging in their own right regardless of the ELF user's L1 (Jenkins 2006: 142) or whether they may be ascribed to interlanguage phenomena (Prat Zagrebelsky 2004).

2. Universal features of translation

Within the descriptive tradition of translation studies (DTS) the whole issue of what constitutes an original text has been called into question on the basis of empirical research within a historical perspective. By investigating the circumstances surrounding the way translations are commissioned, realized and disseminated within their historical and cultural milieu, DTS aims to provide an account of the phenomenon of translation and of the nature of translated text (Snell-Hornby 2006). The approach is largely target-oriented since it endorses Toury's view (1984) that a target text's identity as a translation is determined first and foremost by elements in the receptor system, without there being necessarily any connection with the source text.

DTS has therefore been instrumental in considering translations as autonomous texts which enjoy an independent status in the receiving culture and constitute a genre of their own. This view has given way to the conviction that translations are characterized by certain types of linguistic behaviour that are either unique to, or occur with significantly higher or lower frequency in, translated texts as compared to monolingual text production. Traditionally, translation-related phenomena of this kind have been attributed to source language interference. The source language tends to interfere with the language of the target text at various levels so that "phenomena pertaining to the make-up of the source text tend to be transferred to the target text" (Toury 1995: 275) or, to use Gellerstam's metaphor, the original language leaves its "fingerprints in translation" (2005: 202).

However, its direct contact with a source language is not the only factor which makes the language of translations different from the language of untranslated texts. Translational language differs from non-translational language in many other ways which do not appear to be related to the languages involved in the translation process. Rather, these differences seem to be generated by the process of mediation that takes place in translation. As Baker found in her computer-based corpora research, the translated text possesses specific linguistic features which arise from its nature as "a mediated communicative event" (1993: 243). Mediating between two different texts, languages and cultures, translators tend to adopt similar linguistic features, whatever the languages involved. Since the mid-nineties, these hypotheses on universal features of translation, which had already been put forward in the 80s,[1] have increasingly been identified in translational corpora, parallel as well as monolingual comparable corpora, such as the TEC (Translational English Corpus), developed at UMIST, and a number of abstract categories of translational patterns of behaviour have been observed. Baker (1996) identifies four main categories: 'explicitation', which refers to the tendency of translators to 'spell things out' rather than leaving them implicit, including the practice of adding background information; 'simplification', which is the tendency on the part of translators to subconsciously simplify the language or message or both in the target text; 'normalization', which denotes the tendency to conform to and even exaggerate patterns and practices which are typical of the target language; and, lastly, 'levelling out', which designates the tendency of the translated text to gravitate towards the centre of a continuum of written and spoken modes and to shift away from the two extremes.

What is interesting for our study is that these universal features of translational behaviour are not instances of deviant or substandard translationese but rather forms that are unique to translated text. They are a compromise between patterns of the source language and those of the target language, and, as such, occupy "a third space" of

[1] The first hypotheses on general linguistic properties of translated language concerned a tendency towards explicitation (see Toury 1980; Blum-Kulka 1986), simplification (Blum-Kulka and Levenston 1983; Vanderauwera 1985) and growing standardization (Toury 1980; Vanderauwera 1985).

their own, rather like the characteristics of ELF described by Jenkins (2006: 137). The language used is not manifestly 'foreign' but 'more native than native' in the sense that it aims at normalization, standardization and transparency. There is, besides, a tendency to 'play safe' and thus to use language that conforms to acceptable and conventional linguistic practice in the target culture. The search for cross-linguistic regularities in translational phenomena, as distinct from specific language-pair interference, is mirrored by similar observation-based corpus research in the field of ELF, which is also producing valuable insights into the nature of mediated discourse.

3. Universal features of ELF

Being "a 'contact language' between persons who share neither a common native tongue nor a common (national) culture, and for whom English is the chosen foreign language of communication" (Firth 1996: 240), ELF inevitably absorbs elements from the various linguistic and cultural backgrounds of its speakers. However, despite the hybrid nature of ELF interactions, empirical work conducted on ELF corpora such as the Vienna-Oxford International Corpus of English (VOICE) (Seidlhofer 2004a) and the English as a lingua franca in Academic settings (ELFA) corpus (Mauranen 2003) suggests that ELF usage has specific linguistic elements which, in most cases, cannot simply be ascribed to influences of individual speakers' mother tongues. Mauranen (2006: 156), for example, states that "ELFA findings lend support to the perception that lingua franca English has its own specific characteristics". This view is endorsed by Jenkins (2006: 142), who effectively demonstrates that

> lingua franca varieties are emerging in their own right and exhibiting shared features which differ systematically from NS English norms, regardless of the speaker's L1.

Studies into ELF, carried out mainly in oral and general purpose English, concur in describing shared features at various linguistic and textual levels. On the level of pragmatics, Meierkord (1996) and Knapp and Meierkord (2002) have demonstrated that ELF conversations, rather than reflecting the participants' mother tongues' communicative norms, have their own characteristics, such as pausing to allow transitions between conversational phases, choosing safe topics and using politeness strategies such as routine formulaic expressions. At a lexico-grammatical level, research carried out on the VOICE corpus has brought to light 'typical' errors which are commonly made in ELF interactions irrespective of speakers' first languages and levels of proficiency (Seidlhofer 2001). These include dropping the third person present tense *-S*, using the relative pronouns *WHO* and *WHICH* interchangeably, omitting definite and indefinite articles where they are obligatory in native speaker language use (e.g. "our countries have signed agreement about this", Seidlhofer 2001: 16), overusing certain verbs of high semantic generality, such as *DO*, *HAVE*, *MAKE*, *PUT*, *TAKE*, and adding extra redundancy, for instance through prepositions (e.g. "We discussed about …", Seidlhofer 2004b: 9) or through overdoing explicitness ("black colour" rather than "black" or "how long time" rather than "how long", Seidlhofer 2004a, 2004b).

Within the domain of specialized discourse, Johnson and Bartlett (1999) have identified certain characteristics in their description of International Business English that are common to business practitioners from different mother tongue backgrounds, such as the elimination of pre- and postpositions (e.g. "I'll pay the coffee" instead of "I'll pay for the coffee"), the avoidance of passive forms, the use of a simplified tense system and of a simplified sentence structure.

All these regularly-occurring features reveal a general tendency among ELF users to simplify the English language for the sake of mutual intelligibility. For, as Johnson and Bartlett state, "simplifications make the language more transparent and so aid communication between non-native speakers" (1999: 9). Similarly, as ELF studies are increasingly demonstrating (Mauranen 2007), ELF speakers tend to make use of a variety of explicitation strategies in a common effort to prevent linguistic misunderstandings. What interests us here is that these ELF regularities seem to exhibit elements in common with the abstract categories of translation behaviour observed by Baker (1996). Like translational language, ELF usage shares an orientation towards the target audience. Like translators, who tend to adopt a series of domesticating-like strategies to make the translated text more easily accessible to the target culture audience, non-native speakers of English tend to use various accommodating strategies to make the language more intelligible to their interlocutors or readers. Here, the two paths seemingly appear to diverge since most translators will aim for linguistic correctness while for non-native users "mutual accommodation and communication strategies seem to have greater importance for communicative effectiveness than 'correctness' or idiomaticity in native English terms" (Jenkins *et al*. 2001: 16). Yet, it needs to be said that within DTS, as Hermans (1991) has pointed out, 'correctness' is not a static concept but depends on the notions of correctness prevailing in specific socio-cultural and historical contexts. Acceptable translation behaviour is thus a relative concept which can only be established on the basis of the systematic observation of actual translation performance within the system of potential source text – target text relationships and in the light of the prevailing translation conventions at any point in time. Besides, corpus-based studies into translational behaviour (Baker 1996; Laviosa 2000) have detected an overall tendency to exaggerate typical patterns of the target language in an effort to normalize the target text and thus make it more 'domesticated'. The result is, in actual fact, to neutralize the native elements of the language. All in all, therefore, the priority that translators and non-native speakers alike give to the principles of readability and intelligibility is reflected, as we shall see below, in a common tendency to simplify the source language or message and to make it more explicit.

4. The EUROCOM corpus

To test the existence of similarities between the various forms of mediated discourse, namely the language used by non-native speakers and translational language, we have compiled a corpus of English texts produced within the European Commission (the EUROCOM corpus), which at present comprises about 1 million words. It is a monolingual parallel corpus made up of a non-native and a native component: the non-native set contains original documents written in English by speakers of various mother tongues, who use English as a lingua franca, and the other set is made up of the same documents revised by English native speakers of the European Commission Editing Service. The typology of texts that makes up the corpus is heterogeneous and reflects those the editing service actually has to deal with, ranging from core legislative documents, such as directives, regulations or decisions, to more marginal texts such as speeches by Commissioners, sensitive correspondence, scientific or economic reports, internal administrative matters and staff information. Therefore, highly specialist documents addressed to EU insiders or outside specialists like departmental memos or papers for specialist committees coexist with texts for the general public such as web pages or promotional material. The general purpose of the corpus design is to study the characteristics of written mediated English at various discourse levels, by focusing on the revisions of the native-speaker editors. The revisions are divided into two main categories: obligatory and optional. While the former category mainly consists of basic linguistic errors, the latter includes both personal or stylistic corrections and features which are perceived as strange or non-native by native speakers.

The EUROCOM corpus serves a three-way purpose. Firstly, as the corpus is made up of written documents of English used as a lingua franca in a number of specialized domains, it provides a broader understanding of ELF, whose description has so far been mainly based on spoken corpora of general purpose English, such as the VOICE and the corpus used by Meierkord.[2] Secondly, it allows us to investigate the mediated nature of the non-native versions and assess whether the regularities of non-native English highlighted by the editors' corrections share aspects in common with the tendencies generally related to translational phenomena. Thirdly, it enables the edited versions to be considered as texts in their own right as representatives of intralingual rewriting and thus of mediated discourse. The ultimate aim is to verify whether non-native English elements can also be discerned in the edited texts, which would confirm our working hypothesis that mediation-

[2] ELFA is a corpus of lingua franca English for academic purpose, therefore it is a domain-specific corpus but it restricts itself to the spoken mode. For the reasons why research on ELF has largely been concerned with spoken language, see Mauranen (2006).

specific patterns of language behaviour exist irrespective of the native languages involved and of the nature of the mediation process. It would also lend support to the view that non-native ELF elements are influencing native forms of English.

The results of our initial studies on the EUROCOM corpus indicate that certain salient lexico-grammatical features identified by Seidlhofer in the spoken language can indeed be observed also in the written mode. For instance, one of the most frequent interventions made in the edited documents is the addition of the definite and indefinite articles, which confirms the tendency, clearly deviating from standard English, of non-native speakers to omit the article. It is frequent, for example, to find in the non-native subcorpus sentences like “This document contains following topics” (revised as “This document contains **the** following topics”) or “It should allow for more pro-active policy from MS” (revised as “They should allow for **a** more pro-active policy on the part of MS”). The comparison of the two sub-corpora also reveals certain features identified by Johnson and Bartlett (1999), such as the elimination of pre- and postpositions. Evidence shows that in many cases the editors need to add prepositions after certain verbs or nouns, e.g. *ON* is added after *COMMENT* (as in “83 respondents commented **on** at least one of the R&D and innovation issues”) and after the noun *FOCUS* (as in “The specific **focus** would be ***on*** promoting”), *UPON* is added after *TOUCH* (as in “Half of the respondents **touched upon** this topic”).

As to the identification of translation-related patterns, the editors’ revisions suggest that there are certain common tendencies that non-native authors share with translators, such as a tendency towards increased explicitation, which is expressed both syntactically and lexically. From a syntactic point of view, our quantitative and qualitative studies comparing the edited versions with their originals show that native-speaking editors tend to substitute postmodification by prepositional phrases with premodification by nouns. Thus, in the editing process o*f*-phrases like “officials of the European Commission” are frequently transformed into premodifying nouns, i.e. “European Commission officials”. This type of revision highlights the preference of non-native speakers for constructions which make the information more explicit and treat it as unshared (see Taylor Torsello 1987, Halliday 1993). From a lexical point of view, the initial studies conducted on the editing interventions show the tendency of the editors to delete from the non-native versions certain occurrences of explanatory adverbs such as *MOREOVER* or *IN ADDITION TO THAT*, used to introduce new information, and lexical items which add redundant information, such as *THE EXISTENCE OF* (e.g. "…taking into account factors such as the sector in which a company operates and ~~the existence of~~ any past record of infringement".), *LACK OF* (e.g. "food security […] depends in most cases on ~~lack of~~ access to food") or *AIMED AT CONTRIBUTING* (e.g. "aimed at contributing to the improvement of food security" is rewritten as "for improving food security"). Similarly, in many cases adjectives such as *EUROPEAN* premodifying the name of a European institution, e.g the Parliament, are deleted in the documents addressed to that institution, as the information is already present in the contextual background.

Apart from highlighting the linguistic features that are specific to non-native speakers of English, the editors' revisions provide a valuable tool for investigating the process of editing as another instance of mediated discourse. The editors of the Directorate-General for Translation (DGT) at the European Commission, whose task is to provide linguistic revision of documents drafted by non-native speakers and to offer advisory services to authors, conform to the European Commission's principles of "writing clearly", outlined in the *English Style Guide* and in the *Fight the Fog* booklet, produced by the EC's DGT. The main aim of the editors is to make the documents easily accessible to their readership, which is increasingly composed of the general public and non-native speakers of English. Besides, as a number of these documents will be the basis for translations into several other EU languages, the editors also have the important role of improving the quality of the originals. The editors' mediation between the non-native writers of the draft texts and the end users of the Commission documents essentially consists in rewriting the text in accordance with basic EU principles and in-house conventions to achieve simplification and a plain English style. What is interesting for our research is to see to what extent their rewriting activity can be compared to the types of rewriting produced by the other language mediators and how far changes in English ensuing from its role as a lingua franca influence their editing activity.

5. Universal features of learner English

Insights into the investigation of the universal features of mediated discourse are also emerging from another rapidly developing field of enquiry, that of computer learner corpus research. At a later stage of analysis, the results yielded by the research on the EUROCOM corpus will be checked against the results provided by the research on learner language, which constitutes another form of mediation. Within our research project learner corpora will act as a sort of litmus paper thanks to their predominantly language-specific nature. The influence of the mother tongue is indeed an inherent attribute of the type of language produced by foreign language learners, or interlanguage, a language which is situated somewhere "along a continuum between the native language on one end and the target language on the other end" (Pravec 2002: 109), and which, therefore, inevitably displays features of the native language. One of the main goals of interlanguage research has been to determine the proportion of non-native target language behaviour which is peculiar to native speakers of a given language and that which is shared by all learners of the language, whatever their mother tongue. The exact role of native language transfer is still an unresolved question, but interesting answers have been provided by studies based on such corpora as the International Corpus of Learner English (ICLE) or the Longman Learners' Corpus (LLC), two corpora of learner English from several mother-tongue backgrounds.

Although corpus-based investigations of learner language from various mother-tongue backgrounds highlight the existence of L1-specific patterns (see Aarts and Granger 1998), they also suggest the possibility of a number of cross-linguistic invariants, such as the underuse of the passive, which has been found to characterize the writing of advanced learners of such different mother-tongue backgrounds as Swedish, Finnish and French (Granger 1998: 14). Similar findings have been reported by Martelli (2007: 36) in relation to the use of collocations. While EFL learners have been found "to produce fewer collocations than native speakers and to overuse a smaller number of collocations, especially if these combinations are very frequent in English", thus indicating a tendency towards simplification, evidence from a number of language-specific investigations seems to indicate that "L1 influence plays a considerable role in generating non-native-like collocations" (Martelli 2007: 37).

Lastly, Granger and Rayson (1998) have demonstrated the potential of automatic profiling for revealing the stylistic characteristics of learner language texts vis-à-vis native language and conclude that learner writing displays many of the stylistic features of spoken English. The speech-like nature of learner English can be seen as another pattern of usage that non-native language shares with translational language, that is the tendency to gravitate towards the centre of a continuum of written and spoken modes and to shift away from the two extremes, which, in translation studies, is commonly referred to as levelling out.

6. Conclusion

The recontexualization of knowledge and information (Baker 2006) through various forms of mediated discourse is becoming increasingly widespread in today's globalized world and the co-existence of a language-specific and a universal approach is undeniably an asset in gaining a deeper insight into its nature and the role played by English. The language used largely reflects, and at the same time is conditioned by, the norms prevailing in the international arena. Investigations into translational behaviour within DTS have led Toury (1995) to identify three types of competing norms: mainstream norms, which dominate the scene; dated norms that have been largely superseded; and avant-garde norms that are hovering in the periphery. At present it appears that making texts more accessible to target audiences through universal strategies such as simplification, normalization, explicitation and levelling out constitutes a mainstream norm in translation practice. Ongoing research is directed at verifying whether these recurring patterns of translational behaviour are indeed universal or whether they may be accounted for by the specific languages in contact.

Our initial findings based on a broader concept of mediation, which includes the drafting of texts by non-native speakers of English and the editing and revision of such texts by native English speakers, have led to interesting insights into the hybrid nature of mediated discourse and support the view that hybrid texts are an intrinsic and universal feature of diverse modes of communication. Evidence from the literature demonstrates that they are "characterized by reduced vocabulary, meanings that tend to be universal, and a reduced inventory of grammatical forms" especially within an EU setting (Trosborg 1997: 151). Data from ELF research further substantiate the hybrid nature of interactions involving language contact particularly as far as English is concerned. Because of its role as a lingua franca English is thus particularly susceptible to hybridity, variability and change through contact with other languages via translation, editing or use by non-native speakers.

Further corpus-based studies from an interdisciplinary perspective, involving areas such as translation studies, contrastive analysis, ELF and second language acquisition research, would undoubtedly shed new light on the issues raised by the present analysis and would, as Gentzler (1993: 199) has so aptly observed, make us "less likely to dismiss that which does not fit into or measure up to our standards, and instead open ourselves to alternative ways of perceiving – in other words, to invite real intra-and intercultural communication".

References

Aarts, J. and S. Granger (1998) "Tag sequences in learner corpora: a key to interlanguage grammar and discourse", in S. Granger (ed.), *Learner English on Computer,* Longman, London/New York, pp. 132-141.

Baker, M. (1993) "Corpus linguistics and translation studies. implications and applications", in M. Baker, G. Francis and E. Tognini Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam/ Philadelphia, pp. 233-250.

Baker, M. (1996) "Corpus-based translation studies: the challenges that lie ahead", in H. Somers (ed.), *Terminology, LSP and Translation*, John Benjamins, Amsterdam/Philadelphia, pp. 175-186.

Baker, M. (2006) "Contextualization in translator- and interpreter-mediated events", *Journal of Pragmatics* 38, pp. 321-337.

Blum-Kulka, S. (1986) "Shifts of cohesion and coherence in translation", in J. House and S. Blum-Kulka (eds), *Interlingual and Intercultural Communication*, Narr, Tübingen, pp. 17-35.

Blum-Kulka, S. and E. Levenston (1983) "Universals of lexical simplification", in C. Faerch and G. Kasper (eds), *Strategies in Interlanguage Communication,* Longman, London/ New York, pp. 119-139.

Chesterman, A. (1993) "From 'is' to 'ought': laws, norms and strategies in translation studies", *Target* 5 (1), pp. 1-20.

Delabastita, D. (1989) "Translation and mass-communication: film and T.V. translation as evidence of cultural dynamics", *Babel* 35 (4), pp. 193-218.

Firth, A. (1996) "The discursive accomplishment of normality. On 'lingua franca' English and conversation analysis", *Journal of Pragmatics* 26 (2), pp. 237–259.

Gellerstam, M. (2005) "Fingerprints in translation", in G. Anderman and M. Rogers (eds), *In and Out of English: For Better, For Worse?*, Multilingual Matters, Clevedon, pp. 201-213.

Gentzler, E. (1993) *Contemporary Translation Theories*, Routledge, London/NewYork.

Granger, S. (ed.) (1998) *Learner English on Computer*, Longman, London/New York.

Granger, S. and P. Rayson (1998) "Automatic profiling of learner texts", in S. Granger (ed.), *Learner English on Computer,* Longman, London/New York, pp. 119-131.

Halliday, M.A.K. (1993) "On the language of physical science", in M.A.K. Halliday and J.R. Martin (eds), *Writing Science*, Falmer, London, pp. 54-68.

Hatim, B. and I. Mason (1990) *Discourse and the Translator*, Longman, London.

Hermans, T. (ed.) (1991) "Translational norms and correct translation", in K. van Leuven Zwart and T. Naaijkens (eds), *Translation Studies: The State of the Art*, Rodopi, Amsterdam and Atlanta, pp. 155-169.

Jakobson, R. (1966) "On linguistic aspects of translation", in A. Brower Reuben (ed.), *On Translation*, Oxford University Press, New York, pp. 232-239.

Jenkins, J. (2006) "Points of view and blind spots: ELF and SLA", *International Journal of Applied Linguistics* 16 (2), pp. 137-161.

Jenkins, J., M. Modiano and B. Seidlhofer (2001) "Euro-English", *English Today* 68, 17 (4), pp. 13-19.

Johnson, C. and C. Bartlett (1999) "International business English – What should we be teaching?", *BESIG Business Issues* 3, pp. 8-10.

Knapp, K. and C. Meierkord (eds) (2002) *Lingua Franca Communication*, Peter Lang, Frankfurt/Main.

Laviosa, S. (2000) "TEC: a resource for studying what is 'in' and 'of' translation", *Across Languages and Cultures* 1 (2), pp. 159-177.

Lefevere, A. (1992) *Translation, Rewriting and the Manipulation of Literary Fame,* Routledge, London and New York.

Martelli, A. (2007) *Lexical Collocations in Learner English: A Corpus-based Approach*, Edizioni dell'Orso, Alessandria.

Mauranen, A. (2003) "Academic English as lingua franca – a corpus approach", *TESOL Quarterly* 37 (3), pp. 513–527.

Mauranen, A. (2006) "A rich domain of ELF – the ELFA corpus of academic discourse", *Nordic Journal of English Studies* 5 (2), pp. 145-159.

Mauranen, A. (2007) "Hybrid voices: English as the Lingua Franca of academics", in K. Flottum, T. Dahl and T. Kinn (eds), *Language and Discipline Perspectives on Academic Discourse*, Cambridge Scholars Press, Cambridge, pp. 243-259.

Meierkord, C. (1996) *Englisch als Medium der Interkulturellen Kommunikation. Untersuchungen zum non-native-/non-native speaker – Diskurs*, Peter Lang, Frankfurt/Main.

Paz, O. (1992) "Translation: literature and letters", in R. Schulte and J. Biguenet (eds), *Theories of Translation*, University of Chicago Press, Chicago, pp. 152-162.

Prat Zagrebelsky, M.T. (2004) *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage*, Edizioni dell'Orso, Alessandria.

Pravec, N. (2002) "Survey of learner corpora", *ICAME Journal* 26, pp. 81-114.

Pym, A. (2007) "On history in formal conceptualizations of translation", *Across Languages and Cultures* 8 (2), pp. 153-166.

Sager, J. (1994) *Language Engineering and Translation*, John Benjamins, Amsterdam/ Philadelphia.

Seidlhofer, B. (2001) "Towards making 'Euro-English' a linguistic reality", *English Today* 68, 17 (4), pp. 14-16.

Seidlhofer, B. (2004a) "Research perspectives on teaching English as a Lingua Franca", *Annual Review of Applied Linguistics* 24, pp. 209–239.

Seidlhofer, B. (2004b) "The VOICE of ELF – English as a Lingua Franca", *What's New*?, pp. 8-9.

Snell-Hornby, M. (2006) *The Turns of Translation Studies. New Paradigms or Shifting Viewpoints?*, John Benjamins, Amsterdam and Philadelphia.

Steiner, G. [1975] (1992) *After Babel. Aspects of Language and Translation*, Oxford University Press, Oxford.

Taylor Torsello, C. (1987) *Shared and Unshared Information in English: Grammar to Texts*, Clesp (Unipress), Padova.

Toury, G. (1980) *In Search of a Theory of Translation*, The Porter Institute for Poetics and Semiotics, Tel Aviv.

Toury, G. (1984) “The notion of ‘native translator’ and translation teaching”, in W. Wilss and G. Thome (eds), *Translation Theory and its Implementation in the Teaching of Translating and Interpreting*, Narr, Tübingen, pp. 105-113.

Toury, G. (1995) *Descriptive Translation Studies and Beyond*, John Benjamins, Amsterdam/Philadelphia.

Trosborg, A. (1997) “Translating hybrid political texts”, in A. Trosborg (ed.), *Text Typology and Translation*, John Benjamins, Amsterdam and Philadelphia, pp. 145-158.

Ulrych, M. (1999) *Focus on the Translator in a Multidisciplinary Perspective*, Unipress, Padova.

Vanderauwera, R. (1985) *Dutch Novels Translated into English: The Transformation of a “Minority” Literature*, Rodopi, Amsterdam.

4. Corpora and Specialized Discourse

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 277-291 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Telling a convincing story: a corpus-assisted analysis of business presentations

Julia Bamford – University of Naples, l'Orientale

1. Introduction

The language of business has in recent times begun to attract considerable attention on the part of applied linguists, particularly in the light of the growing importance of how words are used in such areas as marketing, advertising, business negotiations and general business communication. Although much of this attention has been on written discourse, such as job application letters and sales promotion letters (Bhatia 1993), business faxes (Akar and Louhiala-Salminen 1999), economic forecasts (Bloor and Pindi 1990), CEOs' letters,[1] (Garzone 2003), banks' annual reports (Malavasi 2006), it also includes such hybrid genres as business e-mails (Mulholland 1999; Gimenez 2000) and websites (Salvi, Turnbull and Pontesilli 2007) whose language is often informal and conversational. A forerunner in the discussion of spoken business discourse was the work of Bargiela-Chiappini and Harris (1997) on meetings followed by Poncini's (2004) study of negotiations.

Very little attention has been paid to an increasingly significant aspect of business discourse, that of the oral business presentation, although a notable exception to this is Crawford Camiciottoli (2006) who has looked at on-line corporate earnings calls. This particular form of business presentation utilizes various communi-

[1] CEO or Chief Executive Officer is the principal manager of a (usually) large corporation.

cative media including the computer and the telephone and illustrates the heterogeneity of business discourse. It can be formal/informal, take place in/outside the work place, be written or oral, use various channels (for example telephone, face to face, computer) and exploit various registers and genres. In addition, it consists of the recontextualization of many previous discourses and is goal-directed in that language is used to get things done. It varies according to its communicative purpose: it can be promotional, informative, interactional, directive and persuasive. Furthermore, business texts both spoken and written often display all or some of these functions simultaneously.

2. Business presentations

This paper will examine and analyze business presentations which are a relatively under-explored aspect of business discourse. Unlike many spoken genres, business presentations along with other genres such as political speeches, are a carefully prepared form of discourse since the stakes involved are so high. The presentation itself, although seemingly informal, is often attentively prepared by a team and involves the use of sophisticated multi medial devices.

In her ethnographic study of the texts and discourses produced by a manager in his working day, Louhaila-Salminen (2002) has shown that business discourse often involves the use of sets of genres which are connected to one another. Business presentations too are recontextualizations of previous discourse in which information gathered from various sources (statistics, technical information, promotional literature etc.) are put together to form a convincing picture. In this process of choosing from various genres the final presentation speaker is not necessarily the author or only the single author. Furthermore presentations are typically delivered by a team of speakers, usually the CEO followed by the top financial managers. This illustrates Goffman's (1981) distinction between speaker and author and their participation frameworks or Levinson's (1998) further development of this idea under the rubric of multiple participant roles.

3. Convincing the audience

In this paper I will focus in particular on the strategies used by speakers of business presentations to win over their audience by telling a convincing story. Obviously this involves the use of persuasion which has been analyzed since antiquity and is considered a fundamental function of language in use, despite the fact that it has been the focus of few explicit studies. Persuasion, according to Halmari and Virtanen (2005), can be defined as linguistic behaviour which attempts to either change the thinking or behaviour of listeners, or to strengthen beliefs that already exist. In discussing how persuasion is achieved discursively, Hyland (2005) claims that metadiscourse plays a fundamental role as it promotes appeals to credibility, to rationality and to affect.[2] In this he takes up Aristotle's ideas on the rhetoric of persuasion which include ethos (or appeals to credibility), pathos (appeals to affect), and logos (appeals to rationality).

In order to persuade and convince their audience presenters resort to a variety of rhetorical and linguistic strategies including the use of logos or persuasion through reason. Through using argumentation the speaker brings the audience to accept his/her point of view. Ethos too is a significant component of persuasion in business discourse, given the significance of trust in modern capitalism, because it serves to lend the speaker credibility and legitimacy. The integrity and authority of the speaker as perceived by the audience is crucial in a system where investors make decisions based on information provided by the companies. Honesty and candour are regarded as crucial elements in establishing trust with the help of effective communication.

In business presentations a speaker's credibility is of course very dependant on the company's economic and financial performance but how this is presented and evaluated can be more or less effective and persuasive and, crucially, when there is lackluster performance, the credibility of the speaker in promising better results in the future

[2] Hyland shows that various discursive genres (including CEO's letters) are permeated with various types of textual and interpersonal metadiscourse which help to persuade the reader or listener by using, for example, logical connectors, personal pronouns or hedges and emphatics.

becomes of paramount importance. This credibility has to be established and indeed negotiated during the interaction both for the speaker him/herself and as the representative of the company and the persuasive force of the discourse is tied to his/her ability to win the trust of the audience.

Pathos can be seen to be used to good effect also in presenting a company and its financial results. Appeals to the emotions, declarations of belief through personal testimony and attempts to involve the listener through the use of questions and first and second person pronoun address are all persuasive techniques used by speakers in the corpus. These three persuasive strategies, logos, ethos and pathos, are not mutually exclusive and can be used simultaneously in the same discourse sequence as examples from the corpus will show.

One of the strategies used by corporate presenters to persuade the audience is that of telling a convincing story. According to Ochs (2004) and Thompson (2002) the use of narrative in non literary genres plays an important role in, among other things, the construction of personal identity in various genres and discursive social situations. Tellers of a story help to construct a uniform account of what transpires and why, and seek affiliative moral positions from their interlocutors. This is important in presentations which need to present a consistently positive picture of the company and its performance and endevour to engage with the audience to gain their support. One interesting aspect is that presenters themselves recognize the importance of narratives in convincing their audience as the extract below shows.

(1) That then is the Allied Domecq story for this half-year.

Narratives provide a way of construing events which have been experienced by the narrators and experience is filtered through narrative formats. In particular narrative links events in temporal and causal sequences leading speakers to leap from past experience to its implications for the future. Thus the business presentation, as Bowker has observed (2006), often uses narrative to establish communality with the audience and establish credibility.

One striking aspect of business presentations is how personalised it is – speakers use the personal pronouns *I* and *WE* very frequently, call each other by first name, and take responsibility for what they are saying. This is clearly linked to the current trend towards personalization previously noted in various genres including the political; Duranti (2006) for example discusses personalization in election campaign discourse. Fairclough (2000) identifies two key discursive effects of what he calls new capitalism: marketization and informalization; the latter includes the use in public discourse of language practices more typically associated with everyday life.

> The engineering of informality, friendship and even intimacy entails a crossing of borders between the public and the private, the commercial and the domestic, which is partly constituted by a simulation of the discursive practices of everyday life, conversational discourse. (Fairclough 1996: 7)

Fairclough argues that the engineering of informality has two strands: conversationalization and personalization. The former involves the spread into the public domain of linguistic features generally associated with conversation. It is often associated with personalization by which the speaker tries to construct, or at least simulate the construction of a personal relationship with the audience. This can be seen in various genres, including the business presentation, where the speaker tries to develop rapport with the audience through, among other things, the use of names or naming and personal pronouns such as *I* and *WE* typical of conversation.

4. Materials and methods

This study is based on a small specialised corpus of business presentations of roughly 150,000 words, analyzed using ConcApp software (Greaves: www.edict.com.hk). The presentations derive from a variety of sources: some of the transcripts were downloaded directly from company web sites, others were transcribed from the video available online with the help of the power point slides (also available online), yet others were recorded live and transcribed. The presentations in the corpus were given by a variety of companies in a range of settings, often at events organized by merchant banks for the benefit of potential investors. These companies varied from banking to retailing, telecom, mobile telephone services, software, and included companies from various national origins, even if most of the companies were multinationals operating worldwide. Since the audience was not always visible, I relied on information both from informants and the question and answer sessions where members of the audience sometimes gave their name and affiliation before asking the question; it therefore seems that it consisted mostly of experts in the field, such as institutional investors, hedge funds, merchant bankers and representatives of other financial institutions.

The presentations in the corpus are goal directed, transactional and promotional (Bhatia 1993; McCarthy 1998) where the speakers, usually the top of the management hierarchy (the CEO is the most frequent speaker), present the current performance of the company and aim to convince the audience of the solidity of its future prospects. They are usually spoken by multiple participants, although authorship is not acknowledged. According to my informants, due to their vital importance for the company, they go through various drafts and have several authors.[3]

5. The use of personal pronouns in presentations

In examining the corpus one of the first things that strikes the reader is the consistent use of certain personal pronouns rather than others. This can be linked to the conversationalization of business discourse mentioned above, since *I* and *YOU* are among the most frequent words in spoken discourse – the second and third most frequent respectively (Leech *et al.* 2001),[4] in addition they are typically dialogic. According to Wales (1996:52) these prototypically human referents, although seemingly straightforward, have "a wide variety of social and political roles and stances so that the

[3] Two high ranking employees of multinational corporations were interviewed at length to get an insider's account of the significance of and strategies adopted in making business presentations.

[4] McCarthy and Handford (2004) find that in a corpus of spoken business discourse (Canbec) *I* and *YOU* are the fourth and fifth most frequent words and *WE* is a keyword when compared to conversational data. Although in traditional grammars 3pp (personal pronouns) are taken to be typical, in fact 1pp and 2pp are more frequent as Jakobson (1966) saw clearly even before the advent of large corpora proved it.

interpersonal pronouns themselves are rarely ‘neutral’ in their reference.” A frequent pattern in the business presentations in the corpus is centred on the use of personal pronouns linked to the persuasive function of establishing personal credibility or ethos in addition to its use in convincing through affect. The most frequently used is the *I* pronoun which the concordance lines below illustrate.

```
1 these businesses to, as I was going to say, spend money
2 it is available for us.I touched on the multi-brand
3 and uh I think the point I am really trying to stress
4 and to deliver value. As I say, I cannot promise the
5 as it was this year, but I see constant improvement in
6 high provision business I think you are going to see
7 rates going forward, but I think the far more important
8 Australia: Achievements I thought I would say a few
9 for the last two years. I would envisage it being very
10 terms of margin impact I think we were very clear in
11 ratio. On the economy I have not called for any
11 comfortably. However, I would not want to push beyo
```

A closer examination of the concordance lines shows that many of these examples of I and its collocates are being used metadiscursively; the speaker is intruding him/herself into the discourse in order to assist the listener to follow the presentation. Metadiscourse can be defined as the linguistic manifestation of the speaker in the discourse trying to guide the listener towards a better understanding (Hyland 1999: 5). Some metadiscursive concordance lines have been singled out below; while in line 1 the speaker announces the next speaker by name, in line 2 he/she announces the subsequent discourse. Often this personal metadiscourse is cataphoric, as in examples 1 and 2, while in line 4 it is anaphoric referring back to previous discourse and incidentally helping to simulate and consolidate rapport on the basis of a previous encounter.

Metadiscourse is an effective strategy to convince and persuade the listener because the speaker appears to have the interests of the audience at heart, thus enhancing his credibility and creating affect. However, in genres such as business presentations, where trying to persuade and convince is a primary concern, there is often an element of audience manipulation at play.[5]

```
1 number of ways. And so I'd like to ask Jacob
2 in front of customers. I'd like to show you today an
3 ith customers. Now, as I go through this demo,
4 to us. In December, I outlined our disciplined
5 rough all the numbers, I want to make a few opening
```

Besides being metadiscursive, the 1pp with its various collocates seems to function as a politeness marker where the speaker uses tentativeness to refrain from imposing himself on the listener.

```
1 some of those goals. I'm sort of proceeding in the ot
2 e.business we write. I would hope to see further impr
3 about Corporate, and I would guess it is probably the
```

Another frequent pattern consists of the mental verb *I THINK*, a small sample of which is provided below. Weber (as cited in Biber 1988) has pointed out that the subject of this type of verb is usually the 1pp indicating high ego involvement in cognitive processes. *I THINK* is very typical of conversation and thought to express tentative judgment, uncertainly, or lack of commitment (Simon-Vandenbergen 2000). Other context-specific functions of *I THINK*, have been found in various genre studies, for instance, to signal tentative attitude or authoritative deliberation in radio political interviews (Simon-Vandenbergen 2000) and to express authority and high involvement of the speaker in the issue at hand in argumentative essays by both Swedish learners and British and American native English speakers (Aijmer 2001). In these examples the conversationalization of the discourse can once more been seen to be operative since tentativeness is typical of conversation. In addition the speaker lends his/her authority and credibility to the discourse by signaling involvement.

[5] The audience is fully aware of this and consequently attempts to interpret the discourse by 'reading between the lines' according to my informants. This has been noted by Fairclough (1996) also who warns that pathological consequences can ensue from the spread of the simulation of informality: these include recipients questioning the sincerity of the speaker.

1 To be precise, **I think** the best guidance I can give y

2 that. However, **I think** we owe it to our shareholders

3 we operate. **I think** this year's results show we are

In addition, as noticed in Bamford (2007), the presentations in the corpus have a high frequency of the collocates *I* and *BELIEVE* often reinforced by the emphatic operator *DO,* used to express personal conviction and thus persuade the listener. An emotional element is introduced because the speaker commits him/herself to the truth of the propostion thus showing once more the inextricable nature of the ethos and pathos elements in persuasion.

1 seen our advertising. I **do believe** in aggressive

2 the industry. I really **do believe** when you look

3 sales quality is systems. I **do believe** that. If you

In contrast to *I,* the pp *WE* does not seem to be involved in metadiscourse, moreover mental verbs as colligates seem also to be rather infrequent. On the contrary much of the discourse surrounding *WE* regards getting things done, with the speaker trying to convince the audience of the effectiveness of the company and its ability to perform and maintain its promises. This is evidenced by the collocate DELIVER frequently found in the talk following *WE*. The *WE* pronoun signals that the presenter is speaking on behalf of the firm, in particular the management team and the discourse is direct and certain without any of the tentativeness found in the discourse surrounding *I*. Its collocates are action verbs typical of business such as *DELIVER*, *TARGET*, *OUTPERFORM*, and *LEVERAGE*.[6]

1 talked to you a lot about how **we** have delivered those s

2 ive marketing backing it up. **We** will carefully target

3 global products and services. **We** continue to outperform

4 insight not only into the way **we** run our distribution

5 avid Walkden talked about how **we** are going to preserve

6 He showed that not only have **we** already tightened LTVs

7 r the last 12 months but that **we** will continue to do so

8 eclined by 9 percent in 2002. **We** have also leveraged th

[6] Business discourse is characterised by grammatical and lexical innovation, in fact the nouns *LEVERAGE* and *TARGET* have been grammaticalized as verbs.

9 n these administrative tasks. **We** have also seen a doubl

10 r book with an LTV over 85%. **We** were delighted to say

11 hope that you took away that **we** will never compromise

12 promise on our cost profile. **We** delivered 3% on cost

13 and within our card business, **we** are able to provide

14 Those are the reasons that **we** can and will deliver

15 rs in most of our markets as **we** leverage our global

16 nd started with our formula. **We** will not deviate

6. Painting a convincing picture

In business presentations it is essential to paint as positive a picture as possible. When the company's performance is positive this is achieved by emphasis and the use of hyperbole (Bamford 2007) thus making something good seem even better. The evaluative adjective *GOOD* seems to be used to this effect rather frequently. The concordance lines below show how *GOOD* and *FANTASTIC* are displayed. Frequent collocates of *GOOD* include such positive lexis (in a business context) as *GROWTH, INCREASE, PROGRESS,* as well as the more neutral *YEAR, NEWS* and *RESULTS*.

1 was by any standards a very **good** year for HBOS, but

2 ding. Even so, we achieved **good** volume growth in

3 products.So.overall we had **good** volume growth right

4 ed by high employment and **good** affordability. This

5 All of this will produce **good** growth that is

6 strong profit growth, **good** volume growth,

7 We have again seen very **good** profit growth right

8 Treasury is up 8% with a **good** increase in high quality

9 the board, and there are **good**, sustainable increase

10 results show we are making **good** progress along that

11 by anyone's standards a **good** year for Insurance

12 growth. In short, it is a **good** set of results.

13 s and shareholders? The **good** news is that it is

14 it is not easy. This is **good** news because it certainly

15 The first ingredient is **good** cost control. Through

16 it? Obviously we have **fantastic** products, but

17 continue to build on that **fantastic** starting base

7. Good news and bad news

What the speakers in business presentations – in particular those that deal with financial results – are doing is essentially presenting good and bad news. Obviously bad news involves greater problems for the presenter. Financial regulation imposes strict transparency and thus bad news cannot be kept hidden. Nevertheless the corpus data show that the linguistic and rhetorical strategies employed in revealing it could alter the listener's perception. For example, great emphasis is given to the good news in presentations with the speaker placing his/her positive evaluation in first position (as in example 2) and stress placed on continuity (*SUCCESSIVE* and *CONSTANT*) and the vague but very positive *OVER*.

(2) I am delighted to report our eleventh successive half-year of growth with constant currency earnings and interim dividend per share up by over ten percent

On the other hand bad news is generally presented embedded in good news, as in examples 3 and 4 below, where the bad news in 3 is admitted but in the immediately subsequent slot a positive evaluation (*EXCELLENT*) is inserted with the good news. The bad news is not attributed to the company but to the year while the company takes responsibility for the good news.

(3) as I said at the start, the first half of 2005 has been tough but we have an excellent story to tell with constant currency earnings growth of 13 per cent.

In 4 the speaker prefaces the bad news ("revenue will be slightly lower on mobiles") with good news i.e. "strong revenue growth in 3G products". Thus the listener is lulled by a positive assessment placed in first position in the utterance into thinking that the slightly lower revenues are less important than the revenue growth in 3G products. Moreover the bad news is hedged ("slightly lower"). This also illustrates the importance of fulfilling expectations to establish credibility; because the revenue is going to be lower than expected, the company's forecast is shown to be inaccurate and this fact has to be toned down as much as possible. It is not only that bad news is hedged and embedded to lower its impact, but, more important, the resulting unreliability of previous forecasts has also to be downtoned as much as possible.

(4) Whilst strong revenue growth is expected from 3G enabled data products it is likely that the overall rate of increase in proportionate mobile revenue on an organic basis will be slightly lower than that anticipated for the 2006 financial year due to both progressively higher levels of mobile penetration and a greater impact from changes in termination rates.

8. Conclusion

This paper has examined a small specialized corpus of oral business presentations which aim to persuade the listener and to present a positive picture of the past and future prospects of the company. Speakers use many, complex and well rehearsed strategies, often closely interwoven with each other, to persuade their audience. For reasons of space many of these have not been touched on here. What I have chosen to examine are some of the most frequently occurring of these rhetorical strategies, in particular those that illustrate the personalization and conversationalization noted by Fairclough (1996) and the closely connected use of narratives in presentations. Speakers are telling stories, often in the first person, to show personal involvement in the performance of the company and also both to lend and establish their credibility in the attempt to convince the audience.

In order to investigate both these aspects more closely I have chosen to look at first person pronouns and found that several clear patterns emerge from the concordance lines. The first is that *I* is often associated with metadiscursive attempts to guide the listener, but also with hedges to seem non face threatening to the audience, to show commitment and create rapport with the listeners whose interests often do not coincide with those of the speaker. *WE* pronouns, on the other hand, collocate with actions taken by the company which are often positive and show it in a good light.

Another aspect investigated here is the presentation of both good and bad news, a crucial component of this type of business presentation. Bad news, concealment of which is strictly prohibited by financial regulation legislation, is often to be found embedded in good news in order to minimize the impact.

The methodology used, with concordancing supplemented by ethnographic interviews and the specialized nature of the corpus, has permitted a fine grained analysis of the presentations together with some contextualization of the genre. Using corpora to analyze specialized discourse can help to throw up its characterizing features but also to examine critically the techniques used to convince and persuade.

References

Akar, D. and L. Louhiala-Salminen (1999) "Towards a new genre: a comparative study of business faxes", in F. Bargiela-Chiappini and C. Nickerson, *Writing Business: Genres, Media and Discourses*, Longman, London, pp. 207-226.

Aijmer, K. (2001) "*I think* as a marker of discourse style in argumentative student writing", *Gothenburg Studies in English* 81, pp. 247-257.

Bamford, J. (2007) "Accentuating the positive. Evaluation and persuasive discourse in business presentations"; in J. Bamford and R. Salvi (eds), *Business Discourse: Language at Work,* Aracne Editrice, Roma, pp. 135-155.

Bargiela-Chiappini, F. and S. Harris (1997) *Managing Language: The Discourse of Corporate Meetings*, John Benjamins, Amsterdam/Philadelphia.

Bhatia, V.K. (1993) *Analysing Genre: Language Use in Professional Settings,* Longman, London.

Biber, D. (1988) *Variation across Speech and Writing*, Cambridge University Press, Cambridge.

Bloor, T. and M. Pindi (1990) "Schematic structure in economics forecasts", in T. Dudley-Evans and W. Henderson (eds), *The Language of Economics: The Analysis of Economics Discourse,* ELT Document 134, Modern English Publications in association with the British Council, pp. 55-65.

Bowker, J. (2006) "Referential and affective force in oral business presentations: the role of narration", in J. Bamford and M. Bondi (eds), *Managing Interaction in Professional Discourse. Intercultural and Interdiscoursal Perspectives,* Aracne Editrice, Roma, pp. 58-71.

Crawford Camiciottoli, B. (2006) "Rhetorical strategies of company executives and investment analysts: textual metadiscourse in corporate earnings calls", in V.K. Bhatia and M. Gotti (eds), *Explorations in Specialized Genres*, Peter Lang, Bern, pp.115-133.

Duranti, A. (2006) "Narrating the political self in a campaign for U.S. Congress", *Language in Society* 35, pp. 467-497.

Fairclough, N. (1996) "Border crossings: discourse and social change in contemporary societies", in H. Coleman and L. Cameron (eds), *Change and Language,* Multilingual Matters, Clevedon, pp. 3-17.

Fairclough, N. (1996) "The technologisation of discourse", in C.R. Caldas-Coulthard and M. Coulthard (eds), *Texts and Practices. Readings in Critical Discourse Analysis*, Routledge, London, pp. 71-83.

Fairclough, N. (2000) "Discourse, social theory, and social research: the discourse of welfare reform", *Journal of Sociolinguistics* 4 (2), pp. 163-195.

Garzone, G. (2005) "Letters to shareholders and chairman's statements: textual variability and generic integrity", in P. Gillaerts and M. Gotti (eds), *Genre Variation in Business Letters,* Peter Lang, Bern, pp. 179-204.

Gimenez, J.C. (2000) "Business e-mail communication: some emerging tendencies in register", *English for Specific Purposes* 19 (3), pp. 337-351.

Goffman, E. (1981) *Forms of Talk*, Philadelphia University Press, Philadelphia.

Jakobson, R. (1966) *Saggi di linguistica generale*, Feltrinelli, Milano.

Halmari, H. and Virtanen T. (2005) *Persuasion across Genres*, John Benjamins, Amsterdam.

Hyland, K. (1999) "Disciplinary discourses: writer stance in research articles", in C. Candlin and K. Hyland (eds), *Writing: Texts, Processes and Practices*, Longman, London, pp. 99-121.

Hyland, K. (2005) *Metadiscourse: Exploring Interaction in Writing*, Continuum, London.

Leech, G., P. Rayson and A. Wilson (2001) *Word Frequencies in Written and Spoken English,* Longman, London.

Levinson, S. (1987) "Putting linguistics on a proper footing: explorations in Goffman's concepts of participation", in P. Drew and A. Wootton (eds), *Erving Goffman: Exploring the Interaction Order*, Polity Press, Cambridge, pp. 161-227.

Louhiala-Salminen, L. (2002) "The fly's perspective: discourse in the daily routine of a business manager", *English for Specific Purposes* 21 (2), pp. 211-231.

Malavasi, D. (2006) "Banks' annual reports: an analysis of lexical evaluation across some sections", in J. Bamford and M. Bondi (eds) *Managing Interaction in Professional Discourse. Intercultural and Interdiscoursal Perspectives*, Aracne Editrice, Roma, pp.147-158.

McCarthy, M. (1998) *Spoken Language and Applied Linguistics.* Cambridge University Press, Cambridge.

McCarthy, M. and M. Handford (2004) "'Invisible to us': a preliminary corpus-based study of spoken business English", in U. Connor and T.A. Upton (eds), *Discourse in the Professions,* John Benjamins, Amsterdam.

Mulholland, J. (1999). "E-mail: uses, issues and problems in an institutional setting" in F. Bargiela-Chiappini and C. Nickerson Harris (eds), *Writing Business: Genre, Media and Discourses*, Longman, London, pp. 57-84.

Ochs, E. (2004) "Narrative lessons", in A. Duranti (ed.), *A Companion to Linguistic Anthropology,* Blackwell, Oxford, pp. 269-289.

Poncini, G. (2004) *Discursive Strategies in Multicultural Business Meetings*, Peter Lang, Bern.

Salvi, R., J. Turnbull and A. Pontesilli (2007) "The English of companies on-line: national identity and global culture", in J. Bamford and R. Salvi (eds), *Business Discourse: Language at Work*, Aracne Editrice, Roma, pp. 9-45.

Simon-Vandenbergen, A.M. (2000) "The functions of *I think* in political discourse", *International Journal of Applied Linguistics* 10 (1), pp. 41-63.

Thompson, S. (2002) "As the story unfolds: the uses of narratives in research presentations", in E. Ventola, C. Shalom and S. Thompson (eds), *The Language of Conferencing*, Peter Lang, Bern, pp.147-167.

Wales, K. (1996) *Personal Pronouns in Present-day English*, Cambridge University Press, Cambridge.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 293-309 © 2008 Polimetrica International Scientific Publisher Monza/Italy

"In this article, we focus on…": metadiscourse across disciplines

Marina Bondi and Davide Mazzi
University of Modena and Reggio Emilia

1. Introduction

The concept of metadiscourse has been discussed by a variety of authors for more than twenty years now. In particular, metadiscourse has become central in studies on academic discourse and its genres. Following Vande Kopple and Crismore's (1990) study on the effects of hedges on the learning processes of students reading a science textbook, increasing attention has been paid to hedging in particular (Markkanen and Schröder 1997; Hyland 1998a, 1998b; Mauranen 2004). While a number of studies deal with individual, specific categories – e.g. illocution markers (Bondi 2001) or connectors (Bondi 2004) – others include the whole range of tools (see Bamford and Bondi 2005).

Hyland (1998a, 1998b, 2005) insists on the adoption of a comparative framework for metadiscourse studies, and adopts an extensive classification of the various items that can be grouped under the macro-category (hedges, boosters, code glosses etc.) as devices emphasising writer-reader interaction in academic texts.

Accordingly, most recent research has discussed the use of metadiscourse from both a cross-disciplinary and a cross-linguistic perspective (Fløttum and Rastier 2003; Hyland and Bondi 2006). Dahl (2003, 2004) takes a doubly contrastive approach, by investigating metadiscourse as writer manifestation in three languages (English, French and Norwegian) and three disciplines (Economics, Linguistics and Medicine). Bondi (2004, 2005) analyses the role of metadiscursive practices, by comparing historical with economics abstracts.

The present paper moves from a broad notion of metadiscourse, focusing on ‘locational metatext’ comprising “linguistic elements that refer to the text itself or parts of it” (Dahl 2004: 1811), and “rhetorical metatext” (Dahl 2004: 1812), whereby the writer interacts with the reader by making explicit the rhetorical acts he performs in the argumentative process. The aim of the paper is to explore the frequency and use of metadiscursive patterns involving these elements in English research article openings, by comparing two soft-science disciplines, i.e. economics and history.

More specifically, data will be examined through corpus linguistics tools, in order to focus on disciplinary similarities and differences in terms of common metadiscourse functions. Furthermore, more distinctive aspects will be investigated bearing on authorial presence in the organisation of discourse towards both its content and the readership right from the beginning of research articles. Finally, the overall epistemological configurations revealed by collocational and phraseological patterns will be evaluated.

2. Materials and methods

The analysis will be based on a small corpus of 655 research article openings amounting to 196,255 words altogether. The corpus consists in two sub-corpora: the first one is composed of 375 economics openings (100,172 words) taken from eight specialised journals,[1] whereas the second one comprises 280 history openings (95,683 words) from ten journals.[2]

By openings, we mean the first two paragraphs of research articles including epigraphs, if present. For the compilation of the

[1] *European Economic Review (EER), European Journal of Political Economy (EJPE), International Journal of Industrial Organizations (IJIO), International Review of Economics and Finance (IREF), Journal of Corporate Finance (JCF), Journal of Development Economics (JDE), Journal of Socio-Economics (JSE), The North-American Journal of Economics and Finance (NAJEF).*

[2] *American Historical Review (AHR), American Quarterly (AQ), Gender and History (GH), Historical Research (HR), Journal of European Ideas (JEI), Journal of Interdisciplinary History (JIH), Journal of Medieval History (JMH), Journal of Social History (JSH), Labour History Review (LHR), Studies in History (SH).*

corpus, all research article openings were selected for every issue of each journal in the years 1999 and 2000.

As for methodology, three main steps were followed. Firstly, the ten most frequent metadiscourse items were selected from a keyword list obtained by comparing the whole corpus with the written section of the BNC.[3]

Secondly, the selected items were concordanced, in order to identify any similarities and differences between the two disciplines through collocational and phrasal patterns (Sinclair 1996 and 2004b). From this point of view, reference was made to Sinclair's terminology: collocation as the simple "co-occurrence of words" (Sinclair 2004a: 141); colligation as the "co-occurrence of grammatical phenomena" (Sinclair 2004a: 142); semantic preference as "the restriction of regular co-occurrence to items which share a semantic feature" (Sinclair 2004a: 142); and semantic prosody as a "subtle element of attitudinal, often pragmatic meaning" (Sinclair 2004a: 145) words derive from the wider phrasal patterns they occur within.

Finally, each of the two sub-corpora (the economic and the historical) was separately considered for the purpose of a comparison with the written BNC; the aim of this last step was to integrate the former part of analysis, by examining lexical items that more inherently characterise the two disciplines, in order to learn more about their specific epistemological configurations.

The analysis was meant to be both qualitative and quantitative, in order to provide for an accurate account of the pervasiveness of metadiscourse in the introductory part of research articles, where the author(s) set(s) the scene for the rest of the paper.

3. Results

This section is sub-divided into three main parts. In the first section, the metadiscourse items selected through the keyword list are displayed, and their functional similarities across the two disciplinary corpus sections are discussed. In the second section, emphasis is

[3] For the creation of keyword lists as well as the study of concordances, the linguistic software package WordSmith Tools 3.0 (Scott 1998) was used.

laid on differences with an eye to frequency. In the third section, the set of elements considered is extended so as to include metadiscourse tools retrieved by comparing each discipline with the written BNC, which will allow us to formulate hypotheses about deeper epistemological implications behind the use of metadiscourse in economics *vs* history.

Key metadiscourse items: similarities across disciplines

The keyword list created by comparing the whole corpus with the written BNC enabled us to identify the items displayed in Table 1 as the ten most frequent metadiscourse tools shared by both disciplines:

Metadiscourse tools	Frequency	Metadiscourse tools	Frequency
1) *STUDY*	333	6) *ARTICLE*	91
2) *EVIDENCE*	156	7) *ANALYZE*	75
3) *RESEARCH*	136	8) *ASSUMPTION*	67
4) *FOCUS*	130	9) *HYPOTHESIS*	53
5) *ARGUE*	129	10) *PHENOMENON*	49

Table 1. Key metadiscourse tools (corpus *vs* BNC written) and related frequency.

All items listed were lemmatised, by considering singular and plural forms of nouns, and all inflected forms of verbs. Besides, *STUDY* and *FOCUS* were retained in their nominal and verbal use altogether; therefore, the raw frequency reported for them in Table 1 refers to both, whereas the figure regarding *ANALYZE* includes all occurrences of the double spelling *ANALYZE/ANALYSE*.

The concordance-based analysis carried out for each element hinted at three main similarities concerning the use of metadiscourse in economics and history. First of all, metadiscourse is used both to introduce the purpose of the paper,[4] and to represent the broader research field in the openings of both disciplines.

[4] Bondi (2007) explores the statement of purpose that characterizes article openings and studies lexical variation across Italian- and English-speaking academic cultures.

This is corroborated by items such as *ARTICLE*, whose occurrences correspond to the former use in 62.5% of concordance lines in economics (20/32), and in 83.1% of them in history (49/59). On the contrary, *ARTICLE* is used for the purpose of representing the broader research field in 37.5% of its economics occurrences (12/32) and 16.9% in history (10/49 occurrences). The two uses of *ARTICLE* discussed here are exemplified for the two disciplines in (1) and (2) below:[5]

(1a) In this **article**, we will focus… (Economics, EER)

(1b) In this **article**, I shall focus… (History, JMH)

(2a) Several recent **articles** have established the relevance of Joseph A. Schumpeter's theory… (Economics, JSE);

(2b) In 1974 Sir John Sainty published an **article** in which… (History, HR)

In the second place, it was noted that metadiscourse can be used to express the essential interplay between the author's discourse and reported argumentation. This is most clearly signalled by *ARGUE*, which marks the author's own arguments in 12.2% of its economics occurrences (9/74), and 20% of them in history (11/55). By contrast, *ARGUE* conveys reported arguments in 87.8% of attested economics occurrences (65/74) as well as in 80% of history occurrences (44/55). In (3) and (4) below, *ARGUE* as a signal of authorial and reported argumentation respectively is shown at work:

(3) **I argue** that the long chronology was central to Robinson's thought and vision of history… (History, AHR)

(4) **Roubini and Sachs (1989) argue** that the buildup of public debt […] is due to the inability of weak and divided coalition governments to agree on a fiscal reduction package… (Economics, EER)

[5] The sources of all examples are identified by the acronyms provided in footnotes 1 and 2 above. The use of bold typeface as well as underlining in examples signals our own focus of analysis.

In the third place, metadiscourse appears to be involved in a number of patterns common to both economics and history texts. For instance, *FOCUS* as a verb is often part of wider patterns, in which it is preceded by authorial forms such as *I/ WE/ THIS PAPER* or terms that share a semantic preference of ‘research’, e.g. *STUDIES*, *THEORY* and *RESEARCH*. Furthermore, it is followed by words characterised by the shared semantic preference ‘research question’, e.g. *PROBLEM*, *CONTRAST*, *FUNCTION* or *RELATIONSHIP*. The whole pattern described so far could be summarised as follows:

I / WE / THIS PAPER OR ‘Research’ *+FOCUS ON* + ‘Research question’

The chunk occurs 22/83 times in the economics section of the corpus (26.5%), whereas it is attested 15/47 times in the history section (31.9%). In (5) below, two instances of this use of *FOCUS* are provided:

(5a) This analysis suggests that **we focus on the role** of commercial bankers… (Economics, JCF)

(5b) **Modern attempts** to trace the perception of the female body and female sexuality through the history of Western thought **tend to focus on this contrast**. (History, SH)

This sub-section has highlighted some noteworthy parallels between economics and history openings in terms of metadiscourse function and phraseology. In the next section, the attention shifts to the main differences revealed by the study of concordances of the sample of representative metadiscourse tools anticipated earlier in Table 1.

Key metadiscourse items: differences across disciplines

In spite of the common features associated with metadiscourse across disciplines, the analysis of concordances also disclosed a number of differences. These concern two main levels: frequency on the one hand, epistemology at large on the other.

From a merely quantitative point of view, it is apparent that economists employ metadiscourse more frequently than historians, when it comes to opening a research paper. This holds true for every key metadiscourse item considered in this study, with the exception of *ARTICLE*, whose 32 economics occurrences are almost doubled by the 59 occurrences of the history section.[6] As a result, for instance, the use of *ANALYZE* – the item ranked seventh in Table 1 – to introduce statements about the research field concerns both economics (22/49 times, i.e. 37.3%) and history (9/16 times, i.e. 56.3%). Yet, its distribution over the two corpus sections is definitely uneven: 59 occurrences in economics as against 16 in history.

Moving on to more inherently epistemological/methodological annotations, the discrepancies between economics and history became evident through a closer observation of three key-lemmas in particular, notably *EVIDENCE, HYPOTHESIS* and *ASSUMPTION. EVIDENCE* accounts for the second most frequent metadiscourse element of our list, with 88 entries in economics openings and 68 in history. A comparative study of its occurrences shows that economists tend to categorise evidence systematically through adjectives that analytically describe the type of evidence referred to. *EVIDENCE* is preceded by adjectives like *EMPIRICAL*, *DIRECT* and *ANECDOTAL* 21/88 times (23.8%).

On the contrary, historians seem more keen to evaluate evidence rather than categorising it. *EVIDENCE* colligates with either positively or negatively evaluative adjectives 15/68 times, i.e. 22.1%. The difference between the two patterns is illustrated in (6) and (7) below:

(6a) Indeed, there is lot of **empirical evidence** showing that taxes have significant effects on unemployment. (Economics, EER);

(6b) Jaffe (1986) provides additional **direct evidence** of spillovers… (Economics, IJIO)

(7a) Others have moved from discussing such **flawed evidence** to draw more general conclusions… (History, JIH);

(7b) There is **little evidence** of anti-Jewish violence in early twelfth-century England… (History, JMH)

[6] The figure is better explained by looking at the range of self-labelling nouns used in both subcorpora. As shown by Bondi (2007: 77-78), economists tend to use *PAPER* rather than *ARTICLE* for self-reference and such occurrences are 162.

If we consider *HYPOTHESIS* (42 occurrences in economics *vs* 11 in history), a peculiarity of economics is worth pointing out. Unlike history, for which this is never the case in our corpus, economics appears to ascribe hypotheses a central role, by treating them as interpretive keys to review research in the field. This aspect regards 18/42 occurrences of the lemma in economics openings, i.e. 42.8%. The difference in the use of *HYPOTHESIS* in the two disciplines is illustrated in (8) *vs* (9) below:

(8a) … a number of studies have undertaken **to test the hypothesis** in the long-run… (Economics, IREF);

(8b) … others **reject the hypothesis** that there exists a long-run relationship between exchange rate and price ratio… (Economics, IREF)

(9) … with Thiele's 'partial valence' **hypothesis**, there was a stalemate in benzene theory… (History, SH)

In the two examples reported under (8) above, hypotheses act as the yardsticks by which research in the field is respectively reviewed and evaluated, whereas in (9) the term *HYPOTHESIS* only indicates a piece of scientific research that is conventionally attributed the status of hypotheticality. This is well clarified and confirmed by other occurrences of *HYPOTHESIS* in history openings – see "Darwin's hypothesis of creation" (SH), "Kekulé's 'oscillation hypothesis'" (SH) and "the Cartesian 'method of hypothesis'" (SH).

The specificity of economics compared with history can also be seized through the lemma *ASSUMPTION*. Assumptions tend to act as an explanatory criterion of knowledge in economics openings more often than in historical ones, where they frequently appear as an object of negative evaluation and critique. The function just linked with assumptions in economics is mainly expressed by the reiterated occurrence of the pattern represented below:

> *Literature / the discipline / theory / PROPOSITION / HYPOTHESIS / TEST / cases + BE based on / CONSIST of / START from / DEPART from / RELY upon / BE embodied in + the + ASSUMPTION + that…/ of ...*

On its left-hand side, *ASSUMPTION* collocates first with nouns that share a semantic preference of 'research' close to that noted above for *focus*, and then with verbal forms which indicate the most various research steps resting on the assumptions clarified later. The pattern outlined above occurs 10/53 times, i.e. 18.8%.

In contrast, *ASSUMPTION* turns out to be the target of negative acts of evaluation in history 7/14 times, i.e. in 50% of its occurrences. The two uses of *ASSUMPTION* commented so far are illustrated in (10) and (11) below:

(10) Most formal political science and economics **starts from the basic assumption** that individuals are endowed with exogenous and independent preferences. (Economics, EJPE)

(11) … by highlighting the inability of the analytical categories of traditional labour and working-class history, such as class consciousness, skill, or wage, to incorporate the experience of women workers, they **criticized its implicit masculinist assumptions** (History, GH)

Leaving aside quantitative differences concerning the distribution of metadiscourse in the openings of the two disciplines, this sub-section has stressed epistemological distinctions between them on the basis of the study of three key-terms such as *EVIDENCE, HYPOTHESIS* and *ASSUMPTION*. In the upcoming sub-section, the divergent configurations of economics and history are further investigated, by considering key metadiscourse terms imbued with the specific disciplinary background of each corpus section.

Metadiscourse as a key to distinct disciplinary configurations

In order to formulate valid suggestions as to the disciplinary configurations of economics and history, the scope of analysis was enlarged from the items listed in Table 1 and reviewed above to the metadiscourse tools that act as keys to each corpus section. In treating economics and history openings as two separate corpora in their own right, they were compared with the written BNC, in order to retrieve a list of disciplinary metadiscourse items. The two lists obtained in this way through *WordSmith*'s Keyword function are reported in Table 2 below:

Economics openings key metadiscourse tools	**Frequency**	**History openings key metadiscourse tools**	**Frequency**
1) *STUDY*	196	1) *STUDY*	137
2) ***EFFECT***	184	2) *EVIDENCE*	68
3) ***EXAMPLE***	150	3) *ARTICLE*	59
4) *RESEARCH*	99	4) *ARGUE*	55
5) *EVIDENCE*	88	5) *FOCUS*	47
6) *FOCUS*	83	6) ***CONCEPT***	45
7) *HYPOTHESIS*	83	7) ***DISCUSSION***	39
8) ***FACTOR***	77	8) ***ESSAY***	34
9) ***IMPACT***	62	9) ***PHENOMENON***	21
10) *ANALYZE*	59	10) *PARADIGM*	14

Table 2. List of the ten most frequent key metadiscourse tools in economics (economics openings *vs* BNC written) and history (history openings *vs* BNC written), and related frequency.

The two lists in Table 2 show that the vast majority of the elements extracted from the respective keyword lists coincide with the key metadiscourse items studied so far in the paper. This is surely significant, because it confirms the centrality of the analysis completed until now as well as the keyness of the items studied, in spite of the uneven distribution of the latter across corpus sections.

Still, the table emphasises that four items out of the ten listed were not part of Table 1, thus suggesting that they are more intrinsically related to each of the disciplines rather than to the whole of academic discourse represented by historical and economic corpus openings altogether. The items in question are in bold within Table 2: *EFFECT, EXAMPLE, FACTOR* and *IMPACT* for economics; *CONCEPT*, *DISCUSSION, ESSAY* and *PHENOMENON* for history. In the analysis above, they have been taken as a kind of litmus paper reflecting some noteworthy disciplinary concerns of economics and history respectively.

This last part of the paper highlights the pivotal importance taken by the evaluation of cause-effect relationships with regard to empirical facts in economics, as opposed to the centrality of the theorisation and conceptualisation of events in historical debate.

4. The case of economics

The analysis of the four metadiscourse elements listed above for economics, i.e. *EFFECT, EXAMPLE, FACTOR* and *IMPACT* proved particularly interesting for three of them: *EFFECT, FACTOR,* and *IMPACT*. The first two can be grouped together, since they are characterised by a common pattern. They often follow verbs that share a semantic preference of, as it were, 'observation'/ 'scrutiny' e.g. *EXAMINE, INVESTIGATE, ASSESS* and *QUANTIFY*. In addition, they co-occur with the preposition *OF* coming right after them; *OF*, in its turn, precedes nouns whose common semantic preference may be said to be 'economic factors' (e.g. *TAXES*, *ADVERTISING* and *INCOME INEQUALITY*). Finally, these 'factors' are occasionally followed by the preposition *ON* which introduces nouns sharing a semantic preference of 'economic variables' such as *PRICE*, *GROWTH*, *PERFORMANCE* and *UNEMPLOYMENT*. The lengthy pattern described in detail here may be summarised in a more concise way:

> Verbs of 'observation' / 'scrutiny' + *THE EFFECT* + *OF* + 'economic factors' [*ON* + 'economic variables']
>
> Verbs of 'observation' / 'scrutiny' + *THE IMPACT* + *OF* + 'economic factors' [*ON* + 'economic variables']

The whole pattern schematised above for both *EFFECT* and *IMPACT* suggests that economists are particularly interested in evaluating the repercussions of economic factors on the variables that shape the system as a whole. As far as the economics corpus is concerned, *EFFECT* occurs within the above string 39/184 times, i.e. in 21.2% of its occurrences, whereas *impact* occurs in this pattern 13/62 times, i.e. 20.9%.

Interestingly, it was observed that *EFFECT* is attested in the pattern in question in both its singular and its plural form, as indicated by the small capitals used to refer to it as a lemma. On the contrary, only the singular form of *impact* was found when the chunk reproduced above was noted. In (12) and (13), the use of *EFFECT* and *impact* commented here is exemplified:

(12) …we consider **the effect of** changing the number of competitors on market outcome. (Economics, IJIO)

(13) Many studies examine **the impact of** excise or ad valorem taxes on prices, on output of different commodities and on welfare. (Economics, IJIO)

A parallel construct was identified among the concordance lines of *FACTOR*. Data show a considerable degree of post-modification for the lemma, which is frequently followed by *THAT* introducing verbal forms whose dominant semantic preference is ‘causality’ (e.g. *give rise to, influence, trigger, lead to* and *contribute to*). These verbs are then followed by noun phrases denoting ‘economic effects’ such as *high unemployment* or *highly concentrated industrial structures*. The pattern could be condensed into the following formula:

FACTOR + *THAT* + verbs of ‘causality’ + ‘economic effects’

This pattern post-modifies *FACTOR* 23/77 times, i.e. 29.8%, and it is illustrated in (14) below:

(14) the primary purpose of this paper was to model the effect of a state-wide truth-in-sentencing law and to identify the other major socio-economic **factors** <u>that influence the growth of the prison population.</u> (Economics, JSE)

The centrality of the discovery, observation and experimentation of empirical facts as well as the evaluation of their impact and underlying factors is an aspect which appears to cut across the three lemmas considered in this section. In section 5 the focus shifts to history, where, not surprisingly, other disciplinary concerns emerge from the study of relevant metadiscursive lexis.

5. The case of history

As regards the history section of the corpus, *DISCUSSION, CONCEPT* and *PARADIGM* turned out to deserve closer attention. In the first place, data indicate that the discussion of historical dynamics is positively perceived by historians at the outset of their studies. This argument is supported by the fact that *DISCUSSION* is accompanied by an overall positive semantic prosody in 14 of its 39 occurrences, i.e. in 35.9% of them. The discussion of the various objects of the discipline therefore appears as a highly-valued activity within the historical discourse community, either in positive terms (see examples 15a and b below) or in negative terms as in (16), where the lack of consideration by scholars for an important framework of discussion is regretted by the author's voice:

(15a) I shall concentrate mainly on Britain and hope that this will stimulate comparative **discussion**. (History, GH)

(15b) … there has been considerable **discussion** on the class basis of fascism… (History, LHR)

(16) … scholars who have examined Robinson's "new history" have accorded little import to his **discussions** of the vastness of human time. (History, AHR)

The positive emphasis on *DISCUSSION* goes hand in hand with the attempt to conceptualise historical events, their trends and ruptures, namely to reconcile these with super-ordinate reference concepts and interpret them in broader paradigmatic terms.

CONCEPT occurs within constructs that stress its centrality 37/45 times (82.1%). In 22 of those occurrences (48.8% of the total figure), it is followed by *OF* and abstract nouns denoting the disciplinary object discussed in the paper. In the remaining 15 (33.3%), it colligates with positive adjectives which pre-modify it; otherwise, it occurs in co-textual patterns, whereby the use of a concept is underscored as the starting point of a methodological approach, or where the problematicity and complexity of a concept is tackled. In (17) below, examples are provided for the set of formulations upholding the view of the centrality of conceptualisation in history openings:

(17a) This paper focuses on […] the consequences of the introduction of the **concept** of gender on studies of scientific and medical practices (History, GH)

(17b) Thus, the invention of 'The Scientific Revolution', the central **concept** in post-war historiography of science, has frequently been attributed to the Cambridge historian Herbert Butterfield (1900-1979), author of the influential Origins of Modern Science (1949). (History, SH)

(17c) Numerous scholars and countless undergraduates have wrestled with the **concept** of the 'new monarchy' and whether it was initiated in 1485 or found its origins in the polity of Edward IV or earlier. (History, HR)

As for the perception of historical knowledge in terms of events to be set against a sound paradigmatic background, *PARADIGM* deserves to be considered, since it reflects that disciplinary concern in 7/14 occurrences, i.e. 50%. More specifically, the patterns surrounding *PARADIGM* indicate either a reaction to a paradigm or compliance with it 6/7 times altogether, i.e. 85.6% (3/7 occurrences each, namely 42.8%). In the remaining occurrence, the co-textual pattern involving the lemma denotes the establishment of a paradigm. Examples of the three uses of *PARADIGM* described above are provided in (18) below:

(18a) [the] community that has long begun to question the truthfulness and usefulness of the historiographical **paradigm** with whose authorship Butterfield […] has been credited. (History, SH)

(18b) …our fields have been dominated in the postwar period by the same nationalist biases, as well as the same comparative **paradigms** (notably, modernization theory and Marxism) that have prejudiced out European counterparts against provincial places. (History, AHR)

(18c) A unique moment […] came at the turn of the eighteenth century, when the leading chemists of the French Academy of Sciences, […], articulated a Boylean / Cartesian **paradigm**. (History, SH)

In the sections above, some of the distinctive aspects of economics and history in terms of disciplinary configuration have been respectively singled out on the basis of the data taken into account. The conclusion sums up the implications of the findings documented throughout the paper.

6. Conclusion

The paper tried to test the potential of a cross-disciplinary study of metadiscourse in economics and history, by starting from research article openings and studying similarities and differences revealed by the frequency and use of key metadiscourse tools. In particular, it sought to use metadiscourse as a clue to deeply rooted epistemological practices distinguishing economics from history.

At the beginning of the analysis, the focus was on similarities. Data showed that metadiscourse corresponds to two main functional levels in both history and economics. In first place, it introduces the purpose of the paper and/or it represents the broader research field, thus trying to situate the study within the relevant disciplinary debate as well as to set up a dialogue with the respective discourse community. Secondly, it may either signal the author's voice or reported argumentation: from this point of view, metadiscourse tools lay emphasis on the interplay of argumentative voices as a constitutive factor in the construction of disciplinary discourse in openings.[7]

Light was then cast on differences between economics and history. To begin with, discrepancies were noted with regard to frequency, suggesting that economists are more concerned than historians with authorial presence in order to explicitly organise discourse towards both its content and the readership right from the beginning of the paper. In addition, the comparative study of key-lemmas such as *EVIDENCE, HYPOTHESIS* and *ASSUMPTION* enabled us to point to epistemological differences reflecting specific attitudes of the respective discourse community towards the creation and structuration of knowledge in the field.

In sections 4 and 5, this last aspect was further investigated, by extending the list of the items considered to the key metadiscourse elements that more peculiarly denote the two fields. The findings highlighted that at a preliminary stage in research articles, economics appears more centred on the discovery, observation and experimentation of empirical facts and the evaluation of their impact as well as underlying factors. In history, by contrast, objects of the discipline like events, trends and discontinuities tend to be more theoretically discussed, conceptualised and set against the appropriate paradigmatic background.

[7] The polyphonic dimension of academic discourse has been pointed out for history by Bondi and Silver (2006), Bondi (2007) and Bondi and Mazzi (forthcoming).

References

Bamford, J. and M. Bondi (eds) (2005) *Dialogue within Discourse Communities. Metadiscursive Perspectives on Academic Genres*, Niemeyer, Tübingen.

Bondi, M. (2001) "Small corpora and language variation. Reflexivity across genres", in M. Ghadessy, A. Henry and R.L. Roseberry (eds), *Small Corpus Studies and ELT*, John Benjamins, Amsterdam, pp. 135-174.

Bondi, M. (2004) "The discourse function of academic connectors in abstracts", in A.B. Stenström and K. Aijmer (eds), *Discourse Patterns in Spoken and Written Corpora*, John Benjamins, Amsterdam, pp. 139-156.

Bondi, M. (2005) "Metadiscursive practices in academic discourse: variation across genres and disciplines", in J. Bamford and M. Bondi (eds), *Dialogue within Discourse Communities*, Niemeyer, Tübingen, pp. 3-29.

Bondi, M. (2007) "Historical research articles in English and in Italian: a cross-cultural analysis of self-reference in openings", in M. Bertuccelli Papi, G. Cappelli and S. Masi, (eds), *Lexical Complexity: Theoretical Assessment and Translational Perspectives*, PLUS, Pisa, pp. 65-84.

Bondi, M. and D. Mazzi (forthcoming) "The future in history: projecting expectations in historical discourse", in G. Garzone and R. Salvi (eds), *Linguistica, linguaggi specialistici, didattica delle lingue. Studi in onore di Leo Schena,* CISU, Roma, pp. 85-94.

Dahl, T. (2004) "Textual metadiscourse in research articles: a marker of national culture or of academic discipline?", *Journal of Pragmatics* 36, pp.1807-1825.

Hyland, K. (1998a) *Hedging in Scientific Research Articles*, John Benjamins, Amsterdam.

Hyland, K. (1998b) "Boosting, hedging and the negotiation of academic knowledge", *Text* 18 (3), pp.349-82.

Hyland, K. (2005) *Metadiscourse. Exploring Interaction in Writing*, Continuum, London.

Markkanen, R. and H. Schröder (eds) (1997) *Hedging and Discourse. Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*, Walter De Gruyter, Berlin.

Mauranen, A. (2004) "'They're a little bit different…'. Observations on hedges in academic talk", in A.B. Stenström and K. Aijmer (eds), *Discourse Patterns in Spoken and Written Corpora*, John Benjamins, Amsterdam, pp.173-197.

Scott, M. (1998) *WordSmith Tools Manual,* Oxford University Press, Oxford.

Sinclair, J. (1996) "The search for units of meaning", *Textus* IX (1), pp. 75-106; later in J. Sinclair [2004a] *Trust the Text. Language, Corpus and Discourse*, Routledge, London, pp. 24-48.

Sinclair, J. (2004a) *Trust the Text. Language, Corpus and Discourse*, Routledge, London.

Sinclair, J. (2004b) *Reading Concordances*, Longman, London.

Vande Kopple, W.J. and A. Crismore (1990) "Readers' reactions to hedges in a science textbook", *Linguistics and Education* 2, pp. 303-322.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 311-325 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Refreshing the globe? A corpus-based study of 'corporate ethos'

Sandra Campagna – University of Turin

1. Introduction

> Our supplier diversity mission is to provide equal access to procurement opportunities for minority- and women-owned enterprises (MWBEs). We have made a commitment to proactively building relationships with and purchasing goods and services from MWBEs to the maximum extent possible. This mission underscores our long-standing commitment to being a leader in supplier diversity and a model corporate citizen in the communities we serve. In addition, it is keeping with the Coca-Cola Promise "... to benefit and refresh everyone who is touched by our business."

The extract above, taken from the Coca-Cola Company website, highlights, in a nutshell, the guidelines of the company's behavioural code. This entails the promise to foster business development in marginalized communities (and act as "a model corporate citizen" by fulfilling this charitable purpose) as well as the promise to benefit and refresh everyone who is touched by their business, in short the promise to refresh the globe both literally and metaphorically.

If we compare the mission statement quoted above with the following extract taken from the McDonald's website, the similarities are striking:

> The mission of McDonalds Corporation Supplier Diversity is to deliver superior supplier performance through highly qualified minority, women, and small businesses that enhance the overall McDonalds Corporation Customer Experience; support continued economic growth in our diverse communities; and increase global market share.

Both companies promise to respect and comply with local diversity; they both prioritize promoting business initiatives in minority contexts to favour people in need; finally they also share a global view in their commitment to expand the business throughout the planet. How corporate culture and ethical discourse combine to realize the move from global to local (that is, to actualize ‘glocalization’) announced in both texts is the focus of the present paper.

2. Reference to previous studies on ‘glocalization’

‘Glocalization’ is linguistically a lexical blend in that it combines the words ‘globalization’ and ‘localization’. As a market strategy the term indicates “the creation of products or services intended for the global market, but customized to suit the local culture”.[1]

In previous studies I have focussed on the semiotic construction of ‘glocalization’ (realized by definition in terms of ‘moves’ from ‘global’ to ‘local’) in hypertextual representation.[2] In these studies a variety of Web pages (including hypertexts taken from the Coca-Cola website and from the McDonald’s website) has been selected and analyzed within a multimodal framework in line with Kress and van Leeuwen’s seminal work on multimodal representation (Kress and van Leeuwen 1996, 1998, 2001, 2002) and with more recent studies on the dynamic potential of visual communication, particularly suited to the domain of hyperadvertising (Lemke 2002; Iedema 2003; Scollon and Scollon 2003; Janoschka 2004).

For the purpose of the present paper I will only report results relevant to the hypertextual configuration of Web pages taken from the two websites mentioned above. Results emerging from comparing the semiotic construction of ‘glocalization’ on the Coca-Cola Worldwide homepage and on the McDonald’s USA homepage[3] display the following common traits:

- both texts present a triptych structure[4] in which the central

[1] Definition accessibile at: http://www.wordspy.com.

[2] See Campagna (forthcoming).

[3] The Coca-Cola Worldwide homepage is accessible at http://www.cocacola.com/worldwide. The McDonald’s USA homepage is accessible at http://www.mcdonalds.com.

[4] Triptych structures consist of horizontal compositions (to be read from left to right like a printed page) divided into three parts: the first on the left featuring

frame acts as 'The Mediator' between 'Given' and 'New' information;

- the three-frame structure in both texts performs a 'storytelling' function;
- each story realizes 'glocalization' visually.

However, the two homepages construct 'glocalization' differently. Whilst the Coca-Cola Worldwide homepage performs 'glocalization' by enacting a move from 'global' to 'local', the visual story told on the McDonald's USA homepage instead narrativizes 'glocalization' by enacting a move from 'local' to 'global'. Now, using the polarized 'Given/New' information sequence as a main reference point, it follows that the two visual stories present different 'happy endings'. What is newsworthy in the Coca-Cola story, if we apply the 'Given-New' visual polarization framework suggested in multimodal literature, is how the Coca-Cola company realizes the promise to refresh the globe locally, whereas the story told on the McDonald's USA homepage signals a passage from individuality to social aggregation by building up a path of personal growth ranging from self-awareness to community membership.

3. Aims and methods

The present paper has a dual focus. First, it aims at validating/ confuting previous results through the tools of Corpus Linguistics. Second, it aims at expanding on 'glocalization' to verify how business discourse and ethos interrelate in constructing the global/local moves mentioned in Section 2.

To this purpose I adopt the following 5 stage procedure:

1. build two small corpora respectively crawled from the Coca-Cola Company website and from the McDonald's Company website to store quantitative data for comparison;

'Given' information, followed by a central limbo zone acting as 'The Mediator', and the third on the right featuring 'New' information (Kress and van Leeuwen 1996: 217).

2. use WordSmith Tools to obtain frequency lists and concordances;
3. use a larger reference corpus of American English (the FROWN Corpus) to obtain a list of keywords;
4. analyze and interpret data quantitatively/qualitatively;
5. check results against the claims made in the previous studies focussed on the multimodal representation of ‘glocalization’.

4. Using a Web-based Corpus: a few problems

There are a few problems linked both to the ambiguous notion of “Web as Corpus” and to the practicalities that crawling texts from the Web entails.

As to the former, it is worth clarifying that the ubiquitous expression “Web as Corpus” does not have a unified meaning as highlighted in a recent study on Web Corpora (Bernardini *et al.* 2006). Broadly speaking the expression “Web as Corpus” can be used by language researchers either to indicate the Internet as a source of accessible electronic texts or as a corpus proper representing Web English.

As to the latter, the process of constructing Web-based corpora is problematic mainly because of the considerable amount of ‘noise’ produced by Web data which risks distorting results:

> Equally problematic, in terms of linguistic processing and extraction of linguistic information, is the presence of “boilerplate”, i.e., the linguistically uninteresting material repeated across the pages of a site and typically machine-generated, such as navigation information, copyright notices, advertisement, etc. (Bernardini *et al.* 2006: 20)

It is worth mentioning that attitudes to Web ‘noise’ may vary according to ‘boilerplate’ evaluation. In short, if Web redundancy (in the form of automatically generated non-linguistic material and duplicated documents) is valued as counterproductive for research purposes because it might badly affect data interpretation, a filtering process is recommended especially when the target is a large corpus.

> Considerable computational resources are necessary to host a large scale crawl; the data produced by the crawl have to be "cleaned" (removing pages not in the target language or problematic for other reasons; stripping off html code and "boilerplate", discarding duplicates). (Baroni and Ueyama 2006: 4)

On the other hand, 'boilerplate stripping' may not be considered as fundamental if the research focus is not merely on linguistic samples provided by Web sources but also revolves around Web structural components such as the configuration of HTLM documents. In these cases "boilerplate stripping might be undesirable, as it might destroy the logical structure of a document" (Bernardini *et al*. 2006: 21).

The stance taken in the present study with regard to 'boilerplate stripping' has been mixed in the sense that a certain amount of 'noise filtering' has occurred (mainly regarding deletion of animated elements on Web pages).[5] However, this data cleaning process has been here reduced to the minimum because the promotional nature of the selected Web texts requires that the logical structure of such documents is preserved even in the case of duplicates. Finally, there are other problems in Web-based corpora which are not solely ascribable to the Web domain but are part of the current debate in Corpus Linguistics and deal with difficulties in identifying and assessing meaningful texts in corpora (Hunston 2004).

5. Using corpora

Before describing procedure for the data collection of the two small corpora it is worth mentioning that the necessity to integrate corpus analysis with multimodal theory, as indicated in Section 3, justifies here the choice of opting for a corpus-based approach rather than for a corpus-driven approach, since: "Corpus-based approaches favour the integration of corpus linguistics with existing linguistic theories and descriptive frameworks" (Prat Zagrebelsky 2004: 22).

[5] This deletion has involved primarily chunks of Web texts which are not transferable into texts proper because they are protected and this is frequently the case with animated links.

Keywords and keyness

The two small Web-based corpora crawled from the Coca-Cola Company website and from the McDonald's Company website respectively consist of approximately 40,000 words each. More specifically the Coca-Cola Corpus (henceforth referred to as the CCC) totalizes 40,041 words whilst the McDonald's Corpus (henceforth referred to as the McDC) totalizes 35,408 words.

A range of keywords contextualizing 'glocalization'and 'ethos' has been selected on the basis of their ethical component and their keyness in the two corpora has been measured using the FROWN reference corpus.[6]

The tables below shows the ranking of the items in question based on their keyness in the two corpora.

Key words	**Rank according to keyness**
communities	24
corporate	28
environmental	30
local	32
governance	33
global	53
diversity	111
responsibility	169
charities	0

Table 1. Rank according to keyness in the CCC.

[6] Since the culture of the two corpora is American, I have used the FROWN (American English) corpus as a larger reference corpus.

Key words	**Rank according to keyness**
responsibility	9
corporate	11
local	26
charities	27
diversity	31
global	32
communities	37
environmental	50
governance	195

Table 2. Rank according to keyness in the McDC.

Preliminary analysis of quantitative data

The data reported above confirm an interest for ‘glocalization’ policies as shown in keyness of the words *GLOBAL* and *LOCAL* in the two corpora. Moreover, keyness data also indicate a general interest for ethical concerns. However, preliminary observations of quantitative data suggest slightly different orientations. These can be summed up as follows:

- although in both corpora the keyword *LOCAL* reveals a higher degree of keyness compared with the word *GLOBAL*, this is particularly noticeable in the CCC where the gap between *LOCAL* and *GLOBAL* is significantly wide;
- as to ‘ethical’ keywords, words like *DIVERSITY*, *CHARITIES* and *RESPONSIBILITY* seem to be particularly prominent in the McDC as data in Table 2 suggest, except for the word *GOVERNANCE* which displays a higher degree of keyness in the CCC.

6. Analysis of concordances

In order to find how the keywords defining ‘glocalization’ and ‘ethos’ behave in context, four ‘staple’ items have been selected from the original list of 9 keywords and searched using the concord option of WordSmith Tools.

The following tables display the number of occurrences of each item in both the CCC and the McDC and the rate of ethical collocates per ‘staple’ item in the two corpora.

Items	**the CCC**	**the McDC**
local	94	95
global	44	54
community	63	66
corporate	78	103

Table 3. Occurrences of ‘staple’ items in the CCC and in the McDC.

Items	**Ethical collocates in the CCC**	**Ethical collocates in the McDC**
local	50%	47%
global	34%	59%
community	92%	86%
corporate	86%	81%

Table 4. Rate of ‘ethical’ collocates per ‘staple’ item in the CCC and in the McDC.

The quantitative data above do not appear to outline major differences in the two corpora.

As shown in Table 4 both corpora display a high degree of ‘ethical’ collocates especially in relation to the ‘community’ and ‘corporate’ contexts. However, the rate of ‘ethical’ collocates related to ‘global’ indicates a higher degree of ‘ethical’ collocations within the ‘global’ context in the McDC.

7. Selected examples of ‘ethical’ concordances in the CCC and in the McDC

Given the high rate of ‘ethical’ collocates in relation to the ‘community’ context in both corpora, the following examples of concordances displaying ‘ethical concern’ collocates have been selected for qualitative analysis. Following are some concordances for ‘communities’ in the CCC:

1 … As a corporate citizen of these **communities**, we’re affected as well-not …

2 … concerns about health and wellness. ? In **Communities:** we partner with communities …

3 … and our commitment to our consumers and **communities** is great. Our bottling part …

4 … have been largely in tsunami-affected **communities** and have focused on the …

5 … educate parents, children, schools, and **communities** about the critical roles …

6 … and libraries to children from remote **communities** so that they can go to school …

7 … and a model corporate citizen in the **communities** we serve. In addition, it …

8 … to the lives and livelihoods of those **communities.** And we are intensely commit …

9 … our neighbours to help build stronger **communities** and enhance individual …

10 … of 10 more jobs are supported in local **communities.** The most recent study shows …

Following are some concordances for ‘communities’ in the McDC:

1 Given our close relationship with local **communities** around the world, we believe …

```
2 … local business units give back to their **communities** in a wide variety of ways

3 … a net benefit for employees, their **communities,** biodiversity, and the environment …

4 … continued economic growth in our diverse **communities;** and increase global market …

5 … the diversity with our supplier **community** through growing our existing …

6 … just a family restaurant in your local **community**" and your right. The majority …

7 … around the world. McDonald's in the **community** means: Local employment opportunity …

8 Schools for children in impoverished **communities.** McDonald's Brazil engage …

9 … and charitable programs within their **communities.** Requests for local donation …

10 McDonald's UK has collaborated with local **community** groups on anti-litter campaign …
```

Analysis of qualitative data

From the examples above the following ethical areas can be identified in both corpora:

- health and wellness;
- help for destitute populations;
- education;
- environmental concerns.

The 'ethical' component is conceptualized in terms of commitment,[7] which also performs agency attribution as shown in the following examples taken from the concordances listed above:

[7] Emphasis on social commitment also accounts for the high rate of ethical collocates within the 'corporate' context in both corpora.

1 … events. **As a corporate citizen** of these **communities, we're affected** as well-not … (from CCC)

2 … concerns about health and wellness. ? In **Communities: we partner with** communities … (from CCC)

7 … men around the world. **McDonald's** in the **community means: Local employment** … (from McDC)

8 Schools for children in impoverished **communities. McDonald's Brazil engage** … (from McDC)

Ethical commitment is further reinforced by its close association (in terms of collocates) with positively valued items. This positive association projects a favourable aura, that is, it carries a 'Semantic Prosody' of 'ethically correct behaviour'[8] in accordance with Hunston's view on evaluation:

> Hunston suggests that 'what is good' and 'what is bad' can be defined in terms of goal-achievement. Something that is good helps to achieve a goal, while something that is bad prevents or hinders the achievement of a goal. (Thompson and Hunston 1999: 14).

In the following concordances social responsibility collocates are evaluated positively. This reflects Hunston's 'goal-achievement' pattern.

3 … and **our commitment** to our consumers and **communities** is **great.** Our bottling … (from CCC)

4 … have been **largely** in tsunami-affected **communities** and have focused on the … (from CCC)

7 … and **a model corporate citizen** in the **communities** we serve. In addition, it … (from CCC)

[8] The notion of 'semantic prosody' explores the associative property intrinsic in words as units of meaning and 'labels' these associations in evaluative terms. Sinclair was one of the first linguists who acknowledged the importance of 'semantic prosody' in the meaning-making process. He highlights the leading role that the semantic prosody has to play "in the integration of an item with its surroundings" (Sinclair 1996: 87) and defines this role in functional terms: "It expresses "something close to the 'function' of the item – it shows how the rest of the item is to be interpreted functionally" (Sinclair 1996: 87-88).

8 … to the lives and livelihoods of those **communities. And we are intensely commit** … (from CCC)

9 … our neighbours **to help build stronger communities** and enhance individual … (from CCC)

4 … **continued economic growth** in our diverse **communities**; and **increase global market** … (from McDC)

7 … around the world. McDonald's in the **community** means: **Local employment** … (from McDC)

8 … Schools for children in **impoverished communities. McDonald's Brazil engage** … (from McDC)

10 … McDonald's UK **has collaborated with** local **community** groups on anti-litter campaign … (from McDC)

Concordance 8 (from the McDC) is an interesting example of positive evaluation induced by the 'goal-achievement' evaluative scheme where the negative value-loaded concept of 'impoverished communities' is contextualized in positive terms because it collocates with 'McDonald's Brazil engagement'.

Although the concordances in the two corpora display similar traits in the construction of ethical responsibility related to the community context, the 'global' item in the McDC appears to be more prominent. This is shown when explicit reference in the McDC is made to the global view of the company and to the broad notion of 'diversity':

1 Given our close relationship with local **communities around the world**, we believe … (from McDC)

2 … local business units give back to their **communities in a wide variety of ways** … (from McDC)

3 … a net benefit for employees, their **communities, biodiversity**, and the environment … (from McDC)

4 … continued economic growth in our **diverse communities**; and increase global market … (from McDC)

5 … the **diversity** with our supplier **community** through growing our existing … (from McDC)

```
6 … men around the world. McDonald's in the community
means: Local employment … (from McDC)
```

8. Conclusion

On the basis of the quantitative and qualitative data collected and analyzed in Sections 5, 6 and 7, the following conclusions can be drawn.

Both corpora confirm the claims of 'glocalization' advanced in the previous studies on the multimodal representation of corporate culture given the keyness of *GLOBAL* and *LOCAL* in the CCC and the McDC reported in Tables 1 and 2, Section 5.

Moreover, both corpora display a strong correlation between corporate culture and ethical responsibility as shown in Tables 3 and 4, Section 6, where the occurrences of the selected 'staple' items have been measured in the two corpora (Table 3) together with the rate of 'ethical' collocates per 'staple' item (Table 4). In fact, as shown in Table 4, both corpora display a high degree of 'ethical' collocates especially in relation to the 'community' and 'corporate' contexts.

However, although the 'local' dimension prevails over the 'global' in both corpora, the rate of 'ethical' collocates related to 'global' is higher in the McDC.

The qualitative analysis of the sample of concordances listed in Section 7 confirms the wider focus in the McDC. This broader focus is reflected in the realization of a 'Semantic Prosody of 'ethically correct behaviour'. This is expressed through positively evaluated lexis (in Hunston's terms) enhancing the 'global' view of the company and in the wide range of services offered to diverse communities.

These results show consistency with the conclusions drawn from previous studies on the multimodal construction of 'glocalization' (mentioned in Section 2) which revealed different pathways (and different 'happy endings') in narrativizing 'glocalization' in the two given websites. The visual move from 'global' to 'local' anticipated on the Coca Cola Worldwide homepage correlates here with the prevailing 'local' dimension closely linked to ethical concerns in the CCC.

Similarly, the visual move from ‘local’ to ‘global’ which identified ‘glocalization’ on the McDonald’s USA homepage is here reiterated in the concordances extracted from the McDC. And last but not least the different ‘glocalization’ moves which characterize the two corporate companies are further restated in their corporate promises. Whilst the Coca Cola Company promises “to benefit and refresh everyone who is touched by their business” thus focussing on the effect of this revitilizing action on the ‘local’, the McDonald’s Company defines itself as “a family of local restaurants”, in short as a global community (positively evaluated in Hunston’s terms again) made up of local units.

References

Baroni, M. and M. Ueyama (2006) “Building general-and special-purpose corpora by web crawling”, *Proceedings of the 13th NIJL International Symposium*, Tokyo, pp. 31-40.

Bernardini, S., M. Baroni and S. Evert (2006) “A WaCky introduction”, in M. Baroni and S. Bernardini (eds), *WaCky! Working Papers on the Web as Corpus,* Gedit, Bologna, pp. 9-40.

Campagna, S. (forthcoming) “Going ‘glocal’, multimodally speaking”, *ESP Across Cultures* 4.

Hunston, S. (2004) “Counting the uncountable: problems of identifying evaluation in a text and in a corpus”, in A. Partington, J. Morley and L. Haarman (eds), *Corpora and Discourse*, Peter Lang, Bern, pp. 157-188.

Iedema, R. (2003) “Multimodality, resemiotization: extending the analysis of discourse as multi-semiotic practice”, *Visual Communication* 2 (1), pp. 29-57.

Janoschka, A. (2004) *Web Advertising*, John Benjamins, Amsterdam/Philadelphia.

Kress, G. and T. van Leeuwen (1996) *Reading Images. The Grammar of Visual Design*, Routledge, London, New York.

Kress, G. and T. van Leeuwen (1998) “Front pages: (the critical) analysis of newspaper layout”, in A. Bell and P. Garrett (eds), *Approaches to Media Discourse*, Blackwell, Oxford, pp. 186-219.

Kress, G. and T. van Leeuwen (2001) *Multimodal Discourse*, Arnold Publishers, London.

Kress, G. and T. van Leeuwen (2002) “Colour as a semiotic mode: notes for a grammar of colour”, *Visual Communication* 1 (3),pp. 343-368.

Lemke, J. (2002) “Travels in hypermodality”, *Visual Communication* 1 (3), pp. 299-325.

Prat Zagrebelsky, M.T. (2004) “From corpus linguistics to computer learner corpora”, in M.T. Prat Zagrebelsky (ed.), *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners’ Interlanguage*, Edizioni dell’Orso, Alessandria, pp. 11-60.

Scollon, R. and S. Wong Scollon (2003) *Discourses in Place. Language in the Material World*, Routledge, London/New York.

Sinclair, J. (1996) “The search for units of meaning”, *Textus* IX (1), pp. 75-106.

Thompson G. and S. Hunston (1999) “Evaluation: an introduction”, in S. Hunston and G. Thompson (eds), *Evaluation in Text*, Oxford University Press, Oxford, pp. 1-27.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 327-340 © 2008 Polimetrica International Scientific Publisher Monza/Italy

CADIS – A Corpus of Academic Discourse

Maurizio Gotti – University of Bergamo

1. Introduction

This paper focuses on a corpus of texts for academic communication (CADIS), specially compiled by CERLIS, the research centre on specialised languages based at the University of Bergamo. This corpus has been created as an analytical tool to be used within a research project meant to investigate identity traits present in written texts pertaining to various disciplinary contexts. The study is part of a more general project focusing on the relationship between socioculturally-oriented identity-constructing factors and textual variation in different branches of specialised discourse. The project takes into account the internationalisation of specialised discourse in English, not only in Anglophone countries but wherever discursive practices that first appear in English-speaking environments significantly affect also other languages, creating phenomena of 'globalisation' or 'hybridisation' in institutional and professional settings at a local level.

For an in-depth analysis of variation in intercultural communication, our research unit is investigating a variety of texts produced by scholars and academic institutions in different parts of the world, in order to identify textual variants due to the use of English as a first language, a second language, or a lingua franca within the scientific community. The corpus used for our investigation (CADIS) also comprises some Italian texts for comparative purposes. Our approach is not limited, however, to linguistic evidence but is supplemented – wherever possible – with information gathered directly from the interactants about the communicative events and actors involved, and a reconstruction of the general and specific socio-linguistic context of the texts taken into consideration. All these factors are expected to contribute to the definition of identity variants, and their evaluation and interpretation is being carried out in the light of the latest research insights.

The paper first describes the main aims of the research project and, more specifically, of the work of the unit focusing on academic discourse (Section 2). Section 3 concentrates on the criteria followed in the construction of the specialised digital corpus and discusses aspects pertaining to the collection and classification of texts according to their genre, and the gathering of the ethnographic data. Section 4 reports on the investigations currently being carried out with the use of this corpus, while Section 5 provides some concluding remarks.

2. The research project

Our investigation is part of a wider project[1] focusing on the identification of identity traits typical of different branches of specialised English discourse. Within such domains, the project seeks to investigate to what extent the cultural allegiance of (native or non-native) Anglophone discourse communities to their (linguistic, professional, social, national) reference group is affected by the use of English as a lingua franca of international communication. Indeed, the process of internationalisation of English has strengthened the hegemonic tendencies of this language, with the result that local communities are often marginalised and ‘colonised’, thus preventing an authentic intercultural discourse (Wierzbicka 1991; Clyne 1994; Pauwels 1994; Scollon and Wong Scollon 1995; Canagarajah 1999; Fairclough *et al*. 2007).

In some cases, however, virtuous strategies tend to emphasise participants’ specificities and communicative strategies, thus gradually hybridising identities, in contrast or at least in alternative to Anglocentric textual models. As participants strive to suit the

[1] A research project on Identity and Culture in English Domain-specific Discourse funded by the Italian Ministry of Research and coordinated by the University of Bergamo in collaboration with the Universities of Milan, Naples, Turin and IUSM Rome (http://www.unibg.it/cerlis/projects.htm).

communicative needs of an international audience by adapting their native identities to a common plan that implies a new framework of values and shared behaviour, their specialised discourse is bound to combine local roots as well as international conventions. This process is most evident in domains of use (e.g. academic, technical, scientific and legal communication) where the socialisation/textualisation of knowledge plays a crucial cohesive role. The investigation of such genres (Swales 1990, 2004; Bhatia 1993, 2004; Berkenkotter and Huckin 1995; Gillaerts and Gotti 2005; Bhatia and Gotti 2006) and of their diachronic development is a source of valuable evidence as to the language-culture interface, which is also addressed in several ethnographic and sociolinguistic studies.

In the last two decades, the tension between globalising forces and local cultures has been analysed primarily in academic discourse (Ventola and Mauranen 1996; Jordan 1997; Hyland 2000; Flowerdew 2002) and more recently in legal discourse, through a research project on the integrity of legal genres employed in parallel texts in English and other European and non-European languages (Bhatia *et al.* 2003, 2007). There is also a fair amount of research targeting business communication (Ulijn and Murray 1995; Bargiela-Chiappini and Nickerson 1999; Bargiela-Chiappini and Gotti 2005) and institutional discourse (Boden 1994; Martin and Christie 1997) in inter- or multi-cultural settings. However, despite a few exceptions, there is a lack of research aimed explicitly at the investigation of interculturality. In particular, there is a need to investigate the intercultural plane using evidence based on textual data – whether spoken, written or multimodal – from a perspective that is both broad and focused, with textuality seen as the intersection of communicative practices developed by social groups whose settings and aims deserve closer attention in terms of their (linguistic and textual) negotiation of cultural and identitarian aspects.

Early results from the research carried out by our group (Candlin and Gotti 2004a, 2004b; Poncini 2004; Cortese and Duszak 2005; Gotti 2005; Garzone and Sarangi 2007) indicate that the internationalisation which makes English the preferred choice of code is coupled with textual inconsistencies and ambiguities that advise against straightforward, simplified conclusions: the apparent

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

dominance of 'Anglocentric' models in the domains and specialised discourses considered reveals specific adaptive attitudes and evidence of cultural resistance in the textual strategies that construct identity-shaping differences. Furthermore, the background research by members of our research unit has highlighted the complex pragmatic functions of the texts concerned, which are mostly constructed by deploying strong culturally-connoted values.

Within this wider research plan, the Bergamo unit has chosen to investigate the relationship between socioculturally-oriented identity constructing factors and textual variation in academic discourse, not only in Anglophone countries but wherever institutional and professional settings evolve in a way that transcends the linguistic, cultural and conceptual standards of their local communities. In particular, the focus is on the gradual 'globalisation' or 'hybridisation' of discursive practices first appearing in English-speaking environments but also affecting smaller languages, subject to standardising pressures in their semantic, textual, sociopragmatic and even lexicogrammatical construction.

For some scholars (Duszak 1997; Canagarajah 2002; Kandiah 2005) the considerable success of English in the world of academic research poses a threat not only to the survival and productivity of other languages but also for researchers from non-English-speaking cultures, whose construction/perception of specialised discourse inevitably diverges from dominant Anglo-American model(s). In this sense, Mauranen (1993) claims that weaker academic discourses deserve attention and protection on a par with vanishing ecosystems, while Swales (1997) describes English as a tyrant in the field.

Such aspects interact, especially in the case of English, with the evident pressures of transversal identities independent of local traits, thus complicating the overall picture, with a tendency to discourse merging and hybridisation in an intercultural sense. The phenomenon raises a number of questions which deserve further attention: how do specialised discourse communities structure and maintain their identities in an age of globalisation? How do such discourses reflect and/or resolve the tensions between local and global identities? Is it possible to assume a hierarchy of identity-building factors? What methodologies are suitable to describe textual variation viewed from this perspective?

3. The CADIS corpus

As corpora constitute a remarkable tool for the study of discourse, a specific corpus (CADIS = Corpus of Academic Discourse) has been designed as the core and foundation of this project.[2] Since this corpus is a vital element of our investigation and is expected to remain a major resource for years to come, in designing it we have taken into account a number of parameters (e.g. representativeness, sampling methods, balance) so as to make it compliant with international standards and best-practice guidelines.

In view of an in-depth analysis of variation in intercultural communication, our research unit has selected a range of texts produced by scholars and academic institutions in various parts of the world. To identify textual variants arising from the use of English as a first language, second language, or lingua franca of the scientific community, we have devised a corpus formed by English – and in part Italian – texts for academic communication. Beside including two alternative languages, CADIS represents four different disciplinary areas: Legal studies, Economics, Linguistics and Medicine. For each disciplinary area, four different textual genres were considered: abstracts, book reviews, editorials and research articles. The present structure of the corpus is shown in Table 1.

Disciplinary area	No. of articles	No. of abstracts	No. of book reviews	No. of editorials
Legal studies	100	100	100	100
Economics	100	100	100	100
Applied linguistics	100	100	100	100
Medicine	105	105	100	100

Table 1. The present structure of CADIS.

So far, the English texts have been taken from a total of 23 peer-reviewed journals available by subscription through the University of Bergamo website. Because all the journals selected have a high

[2] A detailed presentation of the corpus is given on its webpage at www.unibg.it/Cerlis.

impact factor, we are confident that the content of our corpus is highly representative of each specialised community from which it originated. The same principle is followed in the sampling of Italian academic texts, which are being selected from the most important journals available in each field.

Indeed, apart from representativeness, the structure of our corpus follows a criterion of balance, through an accurate proportioning of its parts. More specifically, when the corpus is completed 50 texts per genre will have been collected and classified within each disciplinary area, totalling 600 texts per discipline and 600 texts per genre. For each language group – native speakers (NSs) and non-native speakers (NNSs) of English, and native speakers of Italian (ITA) – a total of 800 texts (200 per disciplinary area) will have been included in the corpus. Furthermore, within the NS group, equal representation of different varieties (i.e. UK, US, Canadian and Australian English) will be emphasised. When fully implemented, CADIS will comprise 2,400 academic texts, reaching a total of about 2 million tokens (primarily from digital formats but if necessary also from print), selected and classified by disciplinary area, genre, language, author (i.e. NS/NNS), geographical provenance (US, UK, CAN, AUS), date of publication and source journal.

In order to be included in the corpus, a text must be homogeneous in size, with an average of 12,000 words. The selection of similar size samples is meant to simplify later contrastive research, although it is now widely believed that also texts of varying length can provide sound linguistic insights (Sinclair 2004). The samples collected in CADIS consist of entire documents rather than parts of texts, as the integrity and representativeness of complete genres is far more important than the difficulty of reconciling texts of different dimensions. Because CADIS is a specialised corpus, its underlying factor is homogeneity. With this in mind, we have omitted any text whose author, language, genre and source did not comply with our parameters.

The structural complexity of CADIS reflects its contrastive orientation: it is in fact designed to be internally comparable, so that our researchers can analyse and contrast the chosen texts, not only by disciplinary area, genre, language and culture, but also historically.

This is possible because the corpus will cover a time frame of at least 25 years, from the early 1980s to the present day. In the first phase of the selection process, priority has been given to English texts published over the last six years, but when this phase is completed, earlier academic texts will also be selected and archived.

To ensure that the corpus remains useful and workable long into the future for a wide range of studies, a series of measures have been adopted. First of all, backup copies of .pdf and .txt versions of the material ensure that in the event of a system failure or file corruption the corpus can be restored. The texts are saved also in .txt (plain text) format because most corpus processing tools do not accept binary encoding formats such as .pdf, .rtf or .doc. What is more, these commercial formats may not be supported indefinitely in the future (Wynne 2005), so plain text and Unicode (with or without markup) are highly preferable (Smith 2004).

Each text selected is identified with a simple, user-friendly code that summarizes its principal characteristics and enables researchers to immediately recognize and categorize it. An example of this is >MUL MLJ 90(1) 06 NS: 21<, whose parts stand respectively for the author's last name (MUL = Mulroy), the abbreviated journal title (MLJ = *Modern Language Journal*), volume number (90), issue number (1), year of publication (06 = 2006), language (NS = native-speaker English) and total number of pages (21). Relevant information on each text is stored in a separate Excel file, which forms a useful database for the whole corpus.

4. Current investigations carried out with the use of CADIS

The detailed description of each text selected for inclusion in the corpus has been planned so as to enable researchers in our unit to analyze the most significant macro/microlinguistic variants in terms of identity, evaluation and interpretation in the light of recent linguistic scholarship. More specifically, the data are meant to allow an in-depth analysis of the following aspects:

- genre and macrostructure, with their resulting lexico-grammatical realisations;

- speech acts expressing positive/negative evaluation, both exophoric and metatextual;
- the pragmatic, interpersonal plane of discourse (stance, hedging, politeness);
- evidence of popularisation and/or promotional discourse;
- the function of verbal and lexical modality;
- the degree of background knowledge required (content schemata);
- the correlation with such authorial variables as gender and academic standing.

In the last two years several investigations have been carried out with the use of the CADIS corpus, as part of the research programme presented above. For reasons of space, only a few of them will briefly be described in this paper.

Ulisse Belotti has investigated a subcorpus of abstracts from four leading journals in the field of economics, in order to identify and examine manifestations of identity written by Italian scholars. In particular, he has focused on the authors' use of first-person pronouns and critical lexis to affirm their membership of specific social configurations or sub-groups and thereby orient their audience's behaviour and expectations. He has also taken into consideration the generic characteristics of this type of abstract in terms of rhetorical structures and the linguistic realizations of such structures. His findings indicate that criticism of previous studies, i.e. critical analysis of theories, models, arguments and views, and the use of first-person pronouns reveal clear instances of identity, mainly professional. Moreover, it was shown how abstractors established identity not only through the use of first-person pronouns or evaluative expressions but also by employing abbreviations and intertextual devices. As a continuation of this research, Belotti has examined the research articles themselves, to see if and to what extent the findings from the first analysis were confirmed in longer texts written by the same authors. In particular, his research focuses on the following questions: in what is the rhetorical organisation of research articles in the corpus different from that of the abstracts written by the same authors? How and to

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

what extent are evaluative expressions and abbreviations used as manifestations of identity? How and to what extent are self-citations and self-references used? Are there any culture-dependent elements which help to establish the authors' identity?

Larissa D'Angelo has explored academic discourse with the aim of furthering the existing knowledge of gender variation by analysing book reviews by male and female authors within the discipline of Applied Linguistics. While variations between disciplines in academic discourse have been widely investigated (Hyland 2000; Hyland and Bondi 2006), the influence of gender in academic writing is still largely unexplored. The few existing studies concerning this field (e.g. Kirsch 1993) seem to suggest that men and women prefer different linguistic features when they express themselves in the academia and interact with fellow researchers. However, the differences in writing styles and author stances between genders have seldom been discussed taking into consideration the age, experience and authority of the writer in the field. Analysing a subcorpus of CADIS, D'Angelo takes into consideration the use of inter-active resources (transitions, frame markers, endophoric markers, evidentials, code glosses) as well as interactional resources (hedges, boosters, attitude markers, engagement markers, self mentions) of male and female reviewers, and identifies variants due to gender as well as age, experience and authority in the field. In a subsequent paper D'Angelo has investigated how reviewers of different nationality, within the academic discipline of applied linguistics, deal with positive and negative appraisals towards their peers, by analyzing book reviews written in English and Italian, by native and non-native speakers. By examining a subcorpus of CADIS comprising 100 book reviews written in English by authors of different nationality (to which a collection of 20 book reviews written in Italian has been added), her study analyses how the use of positive and negative formulations, pragmatical-rhetorical choices and face-threatening utterances in book reviews written in Italian and English, by native and non-native speakers, are influenced by their authors' cultural identity, also taking into consideration possible variants due to age, experience and authority in the field.

Davide Giannoni has taken into consideration medical editorials representative of (a) native speaker English, (b) non-native speaker English, (c) native speaker Italian, to investigate their macrostructure and linguistic realisations in quantitative as well as qualitative terms. Their generic framework has been compared with that identified by other scholars. The corpus evidence signals that disciplinary identities in English and Italian medical settings produce different sociocultural norms and expectations: as in the related letters to the editor genre, the linguistic realisations of structural patterns are the result of intercultural styles of thought which have developed historically within the discipline. In a subsequent study, by using various sections of the CADIS corpus in a contrastive way, Giannoni has investigated the use of metaphoric expressions and their ability to amplify emotional response or 'affect' (Martin 2000) in the interlocutor. His analysis of NS English research articles published in peer-reviewed journals from four domains (economics, law, medicine, linguistics) has shown that evaluative metaphors vary considerably across disciplines, in terms of source domain, connotations and polarization, and that they are linked not only to disciplinary proclivities but also to a discipline's metaphoric identity.

The analysis of the medical subcorpus of CADIS has led Stefania Maci to identify an emerging genre, i.e. the Research Letter. This genre is used to report original research, as letters are not expected to duplicate material published or submitted for publication elsewhere. Moreover, research letters considered for publication undergo external peer review and have a structure which is similar to that of the Research Article: Introduction, Methods, Results, and Comment. Maci's analysis of the research letter's main features and of the rhetorical strategies employed by authors to persuade their readers has shown that they reflect the conventional criteria of scientific discourse, and that they aim at the realization of their perlocutionary function by means of an objective and impersonal style. In a subsequent study Maci has described how scientific discourse is organized across the genres of research articles and research letters. Based on the analysis of texts collected in the CADIS corpus, she has described the principal features of these genres and has highlighted the main metadiscursive strategies employed by authors to organize scientific discourse, involve their readers, and signal their own attitude.

The main focus of Michele Sala's investigation is the expression of identity in academic texts dealing with legal subjects. His analysis of a subcorpus of articles on International Law authored by native speakers of English discusses the linguistic, rhetorical and textual features revealing a particular stance on the part of the author which can be linked to a specific cultural or domain-specific identity. Indeed, the biographical footnotes accompanying the authors' names show that contributors can be divided into two main sub-groups: on the one hand, the members of the academia (professors, researchers, lecturers), who usually approach legal matters through academic and argumentative texts; on the other hand, the 'practitioners' of the law (lawyers, directors and members of legal offices and institutions), usually working with normative, performative or prescriptive kinds of texts. On the basis of this double distinction, the study investigates whether and to what extent specific professional identities (i.e., scholars *vs* lawyers) influence academic prose. As an expansion of this study, Sala investigates the issue of professional identity in legal research articles by focussing on inter-actional discourse features like interrogative forms. In particular, his main concern is to assess whether and how forensic language influences academic discourse when it is authored by practitioners of the legal field or by people learning and teaching how to become courtroom experts. In a further study he compares the different rhetorical styles and strategies employed by native and non-native speakers of English and by experts of the Common Law *vs* experts of the Civil Law system in discussing legal subjects. The main focus of his analysis is on whether and how the different philosophies behind the two legal systems – i.e., the adversarial approach *vs* the inquisitorial approach, the principle of precedent *vs* the recourse to the civil code, the primacy of witness examination *vs* the primacy of written norms and abstract principles, the emphasis on precision and clarity *vs* the recourse to a formal register and a specialised vocabulary – influence authorial styles and the choice of argumentative-persuasive strategies, especially in terms of data organization, use of quotations and interactional style.

5. Conclusion

As has been shown, although its structure has not yet been completed, CADIS has already proved to be a very useful resource for the investigation of authentic discourse, and the impact of this new research tool is deeply felt in all the areas of the language analysis carried out by our research unit. The full potential of the corpus available is the result of careful design and accurate construction, following sound methodological foundations and clear choices by the researchers involved. Of course, some changes may become necessary at a later stage, due to evolving needs and the availability of new data-processing tools, but we are confident that our work so far has laid the basis for a large, consistent corpus of academic discourse which it is hoped will prove a valuable resource for present and future research.

References

Bargiela-Chiappini, F. and M. Gotti (eds) (2005) *Asian Business Discourses*, Peter Lang, Bern.

Bargiela-Chiappini, F. and C. Nickerson (eds) (1999) *Writing Business: Genres, Media and Discourses*, Longman, London.

Berkenkotter, C. and T.N. Huckin (1995) *Genre Knowledge in Disciplinary Communities: Cognition, Culture, Power*, Lawrence Erlbaum, Hillsdale, New Jersey.

Bhatia, V.K. (1993) *Analysing Genre. Language Use in Professional Settings*, Longman, London.

Bhatia, V.K. (2004) *Worlds of Written Discourse. A Genre-based Approach*, Continuum, London.

Bhatia, V.K., C.N. Candlin and M. Gotti (eds) (2003) *Legal Discourse in Multilingual and Multicultural Contexts: Arbitration Texts in Europe*, Peter Lang, Bern.

Bhatia, V.K., C.N. Candlin and J. Engberg (eds) (2007) *Legal Discourse across Cultures and Systems*, Hong Kong University Press, Hong Kong.

Bhatia, V.K. and M. Gotti (eds) (2006) *Explorations in Specialized Genres*, Peter Lang, Bern.

Boden, D. (1994) *The Business of Talk. Organization in Action*, Polity Press, Cambridge.

Canagarajah, A.S. (1999) *Resisting Linguistic Imperialism in English Teaching*, Oxford University Press, Oxford.

Canagarajah, A.S. (2002) *A Geopolitics of Academic Writing*, University of Pittsburgh Press, Pittsburgh.

Candlin, C.N. and M. Gotti (eds) (2004a) *Intercultural Aspects of Specialized Communication*, Peter Lang, Bern.

Candlin, C.N. and M. Gotti (eds) (2004b) *Intercultural Discourse in Domain-specific English*, Special issue of *Textus* XVII (1).

Clyne, M. (1994) *Inter-cultural Communication at Work: Cultural Values in Discourse*, Cambridge University Press, Cambridge.

Cortese, G. and A. Duszak (eds) (2005) *Identity, Community, Discourse: English in Intercultural Settings*, Peter Lang, Bern.

Duszak, A. (ed.) (1997) *Culture and Styles of Academic Discourse*, Mouton de Gruyter, Berlin.

Fairclough, N., G. Cortese and P. Ardizzone (eds) (2007) *Discourse and Contemporary Social Change*, Peter Lang, Bern.

Flowerdew, J. (ed.) (2002) *Academic Discourse*, Cambridge University Press, Cambridge.

Garzone, G. and S. Sarangi (eds) (2007) *Discourse, Ideology and Ethics in Specialized Communication*, Peter Lang, Bern.

Gillaerts, P. and M. Gotti (eds) (2005) *Genre Variation in Business Letters*, Peter Lang, Bern.

Gotti, M. (2005) *Investigating Specialized Discourse*, Peter Lang, Bern.

Hyland, K. (2000) *Disciplinary Discourses. Social Interaction in Academic Writing*, Longman, London.

Hyland, K. and M. Bondi (eds) (2006) *Academic Discourse across Disciplines*, Peter Lang, Bern.

Jordan, R.R. (1997) *English for Academic Purposes*, Cambridge University Press, Cambridge.

Kandiah, T. (2005) “Academic writing and global inequality: resistance, betrayal and responsibility in scholarship”, *Language in Society* 34 (1), pp. 117-132.

Kirsch, G. (1993) *Women Writing in the Academy: Audience, Authority, and Transformation*, Southern Illinois University Press, Carbondale/Edwardsville.

Martin, J.R. (2000) “Beyond exchange: appraisal systems in English”, in S. Hunston and G. Thompson (eds), *Evaluation in Text. Authorial Stance and the Construction of Discourse*, Oxford University Press, Oxford, pp. 142-175.

Martin, J. and F. Christie (eds) (1997) *Genre and Institutions*, Cassell, London.

Mauranen, A. (1993) *Cultural Differences in Academic Rhetoric*, Peter Lang, Frankfurt am Main.

Pauwels, A. (ed.) (1994) *Cross-cultural Communication in the Professions*, Special issue of *Multilingua* 13 (1-2).

Poncini, G. (2004) *Discursive Strategies in Multicultural Business Meetings*, Peter Lang, Bern.

Scollon, R. and S. Wong Scollon (1995) *Intercultural Communication: A Discourse Approach*, Blackwell, Oxford.

Sinclair, J. (2004) *Trust the Text: Language, Corpus and Discourse*, Routledge, London.

Smith, A. (2004) *A Companion to Digital Humanities*, Blackwell, Oxford.

Swales, J.M. (1990) *Genre Analysis. English in Academic and Research Settings*, Cambridge University Press, Cambridge.

Swales, J.M. (1997) "English as tyrannosaurus rex", *World Englishes* 16, pp. 373-382.

Swales, J.M. (2004) *Research Genres: Explorations and Applications*, Cambridge University Press, Cambridge.

Ulijn, J. and D. Murray (eds) (1995) *Intercultural Discourse in Business and Technology*, Special issue of *Text* 15 (4).

Ventola, E. and A. Mauranen (eds) (1996) *Academic Writing. Intercultural and Textual Issues*, John Benjamins, Amsterdam.

Wierzbicka, A. (1991) *Cross-cultural Pragmatics. The Semantics of Human Interaction*, Mouton de Gruyter, Berlin.

Wynne, M. (2005) *Developing Linguistic Corpora: A Guide to Good Practice*, Oxbow Books, Oxford.

5. Corpora and English Language Teaching

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 343-354 © 2008 Polimetrica International Scientific Publisher Monza/Italy

It's only human…

Guy Aston – University of Bologna

1. Introduction

Learner corpora have generally been subjected to two main types of study – error analysis, where 'mistakes' are marked up by category and then quantified, and comparative analysis, where the frequency of particular features is compared with the frequency of those features in native-speaker corpora. Like errors, such frequency differences are generally interpreted in negative terms, as is clear from the terms 'overuse' and 'underuse' used to describe them. The implication is that the learner should at all times attempt to conform to native-speaker norms, in relation both to individual occurrences (errors) and to general tendencies (frequencies).

In this paper I want to question this focus on native speaker norms in learner corpus analysis, since it ignores a range of factors which – from a teaching and learning perspective – are I think of equal if not greater importance. Since much learner discourse can be seen as successful in achieving its communicative aims, I want to argue that we need to approach learner corpora with other instruments than red ink, ceasing to evaluate difference in purely negative terms.

2. A corpus analysis of error

Since we are talking about corpora, let me outline the terms of the problem by examining the verb which (etymologically) underlies the lemma *ERROR* in a native-speaker reference corpus, the BNC

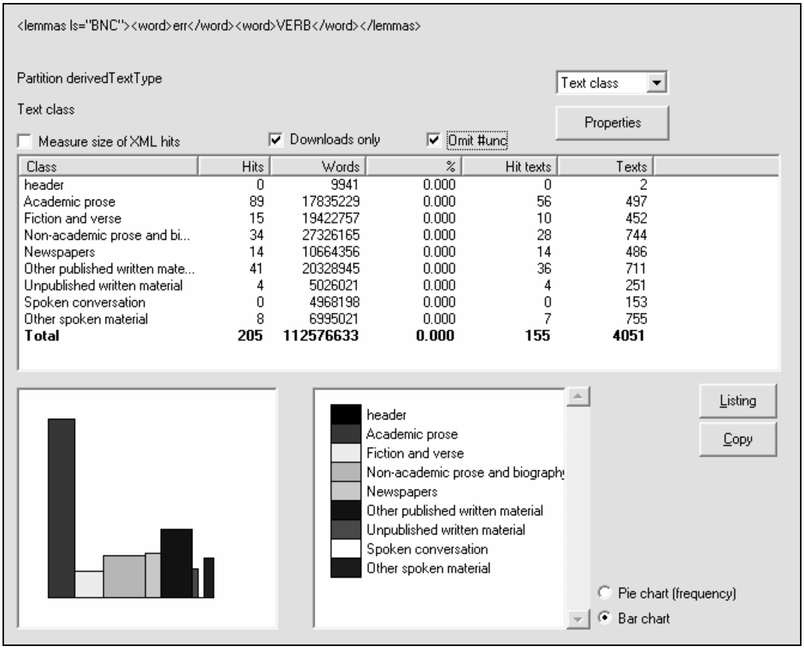
XML edition (BNC-XML 2007). The verb lemma *ERR* is relatively rare in the BNC, occuring only 236 times overall. A few of these are clearly tagging mistakes – the CLAWS C5 tagger appears insufficiently versed in P.G. Wodehouse, idiosyncratic abbreviations, and acronyms:

(1) Bertie: Splendid! (Looks puzzled) Err, what is it Jeeves? (B38: 1653)

(2) Curr Errs (Mandatory Input/Display Field) (HWF: 11243)

(3) Emissions of carbon dioxide and nitrogen oxides from lorries can only be cut by reducing the demand for road freight and not just by improving their fuel consumption, according to the consultants Earth Resources Research (ERR). (J3H: 747)

After removing such spurious hits, along with a handful of misprints, we are left with 205 occurrences from 155 texts. Nearly half of these occurrences (89) come from written academic prose:



<lemmas ls="BNC"><word>err</word><word>VERB</word></lemmas>

Partition derivedTextType

Text class

□ Measure size of XML hits ☑ Downloads only ☑ Omit #unc

Class	Hits	Words	%	Hit texts	Texts
header	0	9941	0.000	0	2
Academic prose	89	17835229	0.000	56	497
Fiction and verse	15	19422757	0.000	10	452
Non-academic prose and bi...	34	27326165	0.000	28	744
Newspapers	14	10664356	0.000	14	486
Other published written mate...	41	20328945	0.000	36	711
Unpublished written material	4	5026021	0.000	4	251
Spoken conversation	0	4968198	0.000	0	153
Other spoken material	8	6995021	0.000	7	755
Total	**205**	**112576633**	**0.000**	**155**	**4051**

○ Pie chart (frequency)
● Bar chart

Figure 1. Distribution of ERR in the BNC.

Let us start by looking at these 89. The majority come from legal texts, where they concern appeals:

(4) the judge had erred in law in holding that the court had no jurisdiction to make any order under section 238 of the Act of 1986 against the bank (FD8: 99)

The collocation *ERR IN LAW* occurs 25 times, where those who err are judges, justices, courts and tribunals. A further 46 cases refer to judicial error where *IN LAW* is left implicit:

(5) the judge had erred in finding that the removal of the child by the mother was not wrongful within the meaning of the Convention on the Civil Aspects of International Child Abduction (FDL: 36)

The dominant use in academic prose (71/89) is thus in reference to errors by judges. Among the remaining 18 cases we find references to errors by experts in other fields:

(6) In the light of this fact it seems possible to suggest that those authors who subscribe to the former view do so because in later times it would have been only in the rarest of circumstances that one would have held the kadilik after the kazaskerlik. It seems likely, in short, that they have erred through applying principles which are valid for a later period but were not yet operative in the time of Molla Husrev. (H7S: 7-8)

(7) Adam is known to be broadly correct in stating that Harald exercised some kind of power in Norway, but obviously errs in having Hiring rule all England. (HXX: 137)

(8) It seems that Beccaria did not often err, in his balancing of contradictory requirements, in favour of the rights of the individual against the achievement of effective deterrence via symbolic representation. (CRX: 99)

(9) Editors, like scientists, are human and do err. (FT3: 1598)

Overall, those who err are experts – the judges, not the judged. In this sense, to err would seem to be a prerogative not of language learners, but rather of those language teachers and corpus analysts who judge their work!

In the academic prose examples we also find a second use of *ERR*, in collocation with *SIDE* – '*ERR* on the side of SUBST' and '*ERR* on the ADJ side':

(10) The therapist should always err on the side of caution; the hypotheses set up are merely shrewd guesses. (EB1: 163)

(11) We should err on the side of restraint, rather than of excess. (GVJ: 570)

(12) Television companies may prefer to err on the safe side rather than to put their profits at risk by incurring sanctions of this order. (J78:401)

What is interesting about this use is the collocates of *ERR* – *CAUTION/THE CAUTIOUS SIDE, SAFETY/THE SAFE SIDE, GENEROSITY/THE GENEROUS SIDE*. When used with *SIDE*, *ERR* appears to have a positive semantic prosody (in Louw's terms: 1993) – to err in this sense is a sensible or praiseworthy strategy:

(13) In general, wherever the officer believes a case to be one which may come to his supervisor's knowledge, it is appropriate to err on the side of caution, to 'go by the book' and protect himself from criticism. (FA1: 1627)

Enough to remind us that learner errors may have a salutary strategic basis!

If we turn to non-academic and miscellaneous published writing, where there are 75 occurrences of *ERR*, we find a not dissimilar picture to that in academic prose. In addition, however, we also find a religious or moral sense, which has heavily negative overtones:

(14) It was a question that troubled Saint Augustine: 'Whence is evil? Why then doth the soul err which God created?' (ACA: 266)

(15) the words of the Anglican matins are made to allude heavily to Grimes's troubles at almost every point: "we have erred and strayed from thy ways" just as Ellen notices the tear in the apprentice's coat (J55: 381)

It is this moral condemnation which seems to have transferred itself to the erring learner in learner corpus analysis. Yet the corpus also reminds us that a learner-centred approach should not morally condemn the one who errs, nor indeed the procedure which generates the errors:

(16) there exists in man's nature an undying capacity to break through the barriers of error, and to seek the road to truth' (*Pacem in Terris* 158). This leads to the distinction between the error (always to be rejected) and the one who errs (always to be respected), and the idea that even 'erroneous' systems can have 'good and commendable elements' (159). (CRK: 637)

Should we not then seek out the "good and commendable elements" in erroneous language use?

3. From *ERR* to ELF: an historical digression

My appeal for a different approach to learner error is not just etymological. I already hear voices protesting that a corpus-based approach cannot assume that *ERROR* has the same uses and meanings as *ERR* (though see Williams' theory of dynamic resonance: Williams 2005). The point I want to make is that in the study of learner error we have in our turn erred – as Pit Corder observed nearly forty years ago:

> I suggest it is misleading to refer to the idiosyncratic sentences of the second language learner as deviant: I also suggest that it is undesirable to call them erroneous as it is to call the sentences of a child erroneous, because it implies wilful or inadvertent breach of rules which, in some sense, ought to be known. (Corder 1971: 18)

Corder's work was seminal in the subsequent development of interlanguage studies. These focussed on the entire production of learners, not just on their errors. An interlanguage was seen as a systematic language variety to be analysed in its own right and with its own dignity, and not just as a collection of unfortunate deviations from native-speaker production. The attitude to learner data which Corder espoused was essentially a neutral one, stressing that the learner's production should not be judged negatively just because it is different. The rise of communicative approaches to language teaching gave a further boost to this change in attitude: it saw a shift in focus from linguistic systems of usage to more pragmatic perspectives of use – how learners do things with words using the foreign language – and began to examine the strategies they (and their interlocutors) adopt to achieve goals of communication and of comity, notwithstanding their limited shared linguistic resources (Aston 1988).

Such a viewpoint provides a focus on success in learner discourse, not merely on failure. It allows us to see error as potentially strategic. And even though much of the literature on interlanguage pragmatics looked at misunderstandings and "communication failure", it also highlighted the extent to which learners and their interlocutors can successfully avoid and overcome such failures, by adopting strategies which are often different from those employed by native speakers. Thus in a survey of this literature, Kasper noted the tendency for learners "to use more transparent, complex, explicit and longer utterances than NS in comparable contexts, and to favour literal over non-literal interpretations" (1997: 350) – in other words, to err on the side of clarity. I recognise this tendency when I write in Italian, where I tend to be more verbose than I am in English, or than an Italian native speaker might be. It's a strategy I consciously adopt to make sure I am understood – erring on the side of comprehensibility at the price of being over-explicit, over-literal, and over-repetitive. But at the same time I try to throw in the odd joke to show I am aware of what I am doing (non-native speakers frequently treat their non-nativeness as an interactional resource in this way, see Aston 1993; Park 2007).

This perspective is not dissimilar from ones in recent ELF (English as a Lingua Franca) studies:

> Some tend to approach the description of ELF data more through the lens of familiar ENL forms, essentially asking, "How do ELF speakers differ from ENL speakers?" [...] Others conceive of ELF differently [...] essentially asking "How do ELF speakers communicate? What seems important/useful to them?" (Seidlhofer *et al.* 2006: 10)

As in interlanguage studies, in other words, we find approaches which focus on the formal characteristics of the linguistic system and its differences from the native-speaker (ENL) system, alongside approaches which focus on its functional characteristics – how learners and ELF users do things with words, and how these differ from native-speaker uses. These differences are not formulated as under/overuse with respect to ENL norms. Rather they are seen as due to differences in usage (interlanguages and ELF are seen as distinct linguistic varieties in their own right), and to differences in the use made of those systems by successful users (ENL and ELF speakers employ distinctive strategies).

What does this mean for the analysis of learner corpora? It implies that we should not just mark up errors, but also communicative successes – the things we put big ticks against when marking learner essays, rather than underlining or crossing out. Similarly, we should be prepared to view differences in frequency from native-speaker discourse as potentially positive factors. While it is a perfectly valid research exercise to compare native and non-native discourse for the purpose of establishing recurrent differences, we need to do so without assuming that the native speaker constitutes a communicative ideal with respect to which non-native performance should be judged. Thus we should jettison such terms as 'error', 'underuse' and 'overuse', which are far too value-laden (Leech 1998).

Let us take some instances from the literature. De Cock *et al.* (1998) found that in speech learners used far fewer 'vagueness tags', such as *AND SO ON*, than native speakers did. Is this relative "underuse" necessarily a bad thing? Vagueness would hardly seem a merit in terms of Grice's maxims (1967) – though it may be one from perspectives of politeness and interaction management. In another study, Martelli (2007) notes how, in comparison with native speaker data, there would appear to be

> a tendency among learners to produce fewer collocations than native speakers and to overuse a small number of collocations, especially if these combinations are very frequent in English or similar in structure to collocational patterns in the L1. (Martelli 2007: 36)

But can we be sure that learners would necessarily do better in communicative terms by producing more numerous and varied collocations, any more than we can be sure they would do better to use more 'vagueness tags'? In his discussion of a study of lexical density by Ringbom (1998), Cobb (2003) makes a very similar point:

> It is plausible that repetition of high frequency items and failure to nuance common notions may well account for the sense of vagueness that native speakers find in advanced learner writing. Admittedly, the evidence is merely correlational: there is vagueness,

> and there is overuse of high frequency lexis, but no causal connection is actually established. The next step in the research agenda is presumably experimental hypothesis testing. Here, for example, Ringbom might have gone on to empirically test teachers' vagueness ratings against learner texts of varying lexical density [...] (Cobb 2003: 400)

Cobb is here calling for the use of subjective ratings by teachers of the communicative effectiveness of the discourse. And if what is at issue is communicative success, should we really assume that native-speaker student essays, such as those used for comparison in the ICLE[1] project, constitute a model to imitate? The error in comparing non-native and native speaker production lies in assuming that the latter is intrinsically better, so that any qualitative or quantitative difference is by definition to be evaluated negatively.

We can only overcome this problem if we evaluate success as well as failure – trying to identify what is effective and right, as well as what is ineffective and wrong. Our teaching experience should surely tell us that focussing solely on error will encourage students to adopt avoidance strategies, and discourage them from taking risks – hardly a good context for learning. Our analyses of learner discourses should show how and where they work, not just how they don't: and what is right is what works in that context, regardless of whether or not it conforms to typical native-speaker performance.

If I may be excused a further comparison *en passant*, I believe we can see a similar difficulty in the area of translation studies. It has been claimed that translations into a given language are systematically different along a number of quantifiable parameters from texts written directly in that language (Baker 1993). As with learner corpora, evaluation is usually implied – the assumption being that translations into a language should resemble original texts in that language, and are to be judged negatively insofar as they differ from them. But again, this assumes that all original texts are perfect, which they clearly are not. If we turn for a moment to consider translation as a real-world process, we can immediately see that it is mainly a matter of editing, with the aim of – if anything – bettering the communicative effectiveness of the original text. If by

[1] International Corpus of Learner English.

any chance a translator takes on the text I have produced here, I sincerely hope that they will make it communicatively more rather than less effective than the original, reducing its vagueness, repetitiveness and general lack of argumentative clarity. And I will only judge it a good translation if they succeed in doing so!

4. A possible moral for learner corpora

To sum up, what is missing in most studies of learner corpora is an evaluation of language use – how pragmatically effective is this text as discourse? Most teachers would, I think, prefer an interesting essay which argues a point in an effective manner, to a mundane one which is free of formal errors. The computer, of course, cannot make such judgments reliably (as users of style-checkers know to their cost). But humans can, and do. I would like to see learner corpora which, as well as marking up orthographic, lexicogrammatical, collocational and stylistic errors in the text, also included metadata regarding the communicative and comitive effectiveness of the discourse – its interest, insightfulness, clarity, coherence, reader-friendliness, seductiveness/annoyingness – along with suggestions for its improvement. Unlike an error-based approach, such reader reactions, being subjective, require analyses of a qualitative nature:

> In the first case, the focus is more on a (quantitative) analysis of forms, whereas [in] the second, it is more on a (qualitative) understanding of processes – allowing, of course, for any amount of gradation and interaction between the two. (Seidlhofer *et al.* 2006: 10)

To arrive at a more quantitative understanding of processes, we will require multiple readers, multiple comments, along the lines suggested by Cobb (2003).

To put it another way. Computer programmes may be able to analyse the formal properties of text,[2] but they do not realise that

[2] Computers can identify much non-standard spelling, some non-standard grammar, and they can count such features as lexical and collocational frequencies, lexical density, sentence and paragraph length. There is little else: even reliable part-of-speech taggers have yet to be developed for learner and ELF corpora. And all these numbers are uninterpretable for significance without an adequate analysis of their variability and distribution in large numbers and varieties of texts.

text as discourse. They do not reproduce the step-by-step interactive negotiation which the reader engages in. Unlike computational analyses, reader's reactions often cannot be matched precisely to particular segments of the text. They involve a different kind of scope – referred not to a static, clearly-identified piece of the text, but to the reader's experience up-to-now in realising that text as discourse, in an interaction which may be smooth or laboured, enjoyable or frustrating. The reader is in the first place a participant in the discourse process.

In the second place, however, the critical reader is also an observer of that process, concerned to establish how her/his reactions are brought about, how communicative success is achieved and why communicative failure occurs. This is not always easy: as an observer, the reader has to make hypotheses as to the writer's communicative intent, in illocutionary and perlocutionary terms, in order to evaluate the discourse process as strategic action. Corder (1971) suggested that translation of the learner text into the source language might help provide contextual information to infer the interlanguage user's intentions. Others have suggested that we should interview the learner in order to understand what was meant and hence to evaluate the strategies employed (MatteBon 2007).

5. Conclusion

I want to thank Micia Prat Zagrebelsky for two brief discussions which prompted me to put pen to paper on these matters (or at any rate finger to keyboard) – and to add that responsability for the limited results here is wholly mine. It has not been my aim to propose a formal schema for marking up and analysing learner corpora, but simply to suggest directions in which I feel we should be moving. Currently, I would argue, we are looking at a tiny (and biased) subset of what we ought to be looking at. We ought to be engaging in and looking at discourse, not just text, and we ought to be looking at success, not just at failure. This implies, I have tried to suggest, a rather different approach to difference and to error – as a final glance at *ERR* in the BNC may serve to remind us:

(17) ‘Anybody can make a mistake. To err, it's only human. I can forgive them for that. (AM4: 597-9)

As well as being human, to err can also be a good idea in terms of learning:

(18) Err towards brevity: you want to leave people wanting more, so that they ask questions. (CEF: 2329)

Consequently, as judges of learners' performance, we need to start taking a more forgiving approach:

(19) The next day was the same. Exactly the same. Except that this time the note said, 'To forgive, divine…' (H8S: 403-5)

In approaching learner corpora, adopting a more forgiving approach may also help us to divine means of analysis which cast greater light on that extraordinary phenomenon of the successful foreign language learner and user. As judges of the learner's performance we will err less by doing so.

References

Aston, G. (1988) *Learning Comity*, Cooperativa Libraria Universitaria Editrice, Bologna.

Aston, G. (1993) "Notes on the interlanguage of comity", in G. Kasper and S. Blum-Kulka (eds), *Interlanguage Pragmatics*, Oxford University Press, New York, pp. 224-250.

Baker, M. (1993) "Corpus linguistics and translation studies – implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam/ Philadelphia, pp. 233-252.

BNC-XML (2007) *The British National Corpus – XML Edition*, Oxford University Computing Services, Oxford.

Cobb, T. (2003) "Analyzing late interlanguage with learner corpora, Quebec replications of three European studies", *Canadian Modern Language Review* 59, pp. 393-423.

Corder, S.P. (1971) "Idiosyncratic dialects and error analysis", in S.P. Corder [1981] *Error Analysis and Interlanguage,* Oxford University Press, Oxford, pp. 14-25.

De Cock, S., S. Granger, G. Leech and T. McEnery (1998) "An automated approach to the phrasicon of EFL learners", in S. Granger (ed.), *Learner English on Computer,* Longman, London/New York, pp. 67-79.

Grice, H.P. (1967) “Logic and conversation”, in P. Cole and J.L. Morgan (eds) [1975], *Syntax and Semantics. Speech Acts,* Vol. 3, Academic Press, New York, pp. 41-58.

Kasper, G. (1997) “Beyond reference”, in G. Kasper and E. Kellerman (eds), *Communication Strategies: Psycholinguistic and Sociolinguistic Perspectives*, Longman, London, pp. 345-360.

Leech, G. (1998) “Learner corpora: what they are and what can be done with them”, in S. Granger (ed.), *Learner English on Computer*, Longman, London/New York, pp. xiv-xx.

Louw, B. (1993) “Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies”, in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair,* John Benjamins, Amsterdam/ Philadelphia, pp. 157-176.

Martelli, A. (2007) *Lexical Collocations in Learner English: A Corpus-based Approach*, Edizioni dell'Orso, Alessandria.

Matte Bon, F. (2007). “Looking at Mellange”. Unpublished contribution to Mellange seminar, SSLMIT Forlì, 25 September 2007.

Park, J.E. (2007) “Co-construction of nonnative speaker identity in cross-cultural interaction”, *Applied Linguistics* 28, pp. 339-360.

Ringbom, H. (1998). “Vocabulary frequencies in advanced learner English: a cross-linguistic approach”, in S. Granger, *Learner English on Computer,* Longman, London/New York, pp. 41-52.

Seidlhofer, B., A. Breiteneder and M.L. Pitzl (2006) “English as a lingua franca in Europe: challenges for applied linguistics”, *Annual Review of Applied Linguistics* 26, pp. 3-34.

Williams, G. (2005) “The Good Lord and his works: a corpus-based study of collocational resonance”, paper presented at *Phraseology 2005*, Louvain-la-Neuve.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 355-364 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Keeping the corpus-based promise to language teaching in schools: the need for a quantum leap

Umberto Capra – University of Piemonte Orientale

1. Introduction: Habeas Corpus

Sinclair (1987: 158) drawing his conclusions about the perspectives opened for language learning by Corpus Linguistics stated that

> The prospect arises of being able to present the facts of the language in a much more precise way than before. Instead of individual words and phrases being crudely associated with a 'meaning', we could see them presented in active and typical contexts, and gradually freed from those contexts to enjoy, in most cases, a severely limited autonomy. Very few common words are thought to have a residue of patterning that can be used independently.

The perspective Sinclair was then referring to was fundamentally a descriptive one: "Each sense of the phrase is co-ordinated with a pattern of choice that helps to distinguish it from other senses." (1987: 158). Yet 'the learner' was mentioned as the main intended recipient of the new evidence offered by corpus analysis: "The learner does not need to be faced with a featureless list of phrasal verbs and no guidance as to which is which and why they are preferred to single words" (1987: 158).
Although warning that

> This material is not intended for direct exploitation in the classroom (though many classes respond well to being offered fairly new data). It is gradually building up as a database for teachers' reference, a repository of facts about English on which new syllabi and materials can be based. (1987:158)

Sinclair already had in mind possible changes in learners' strategies (1987: 159):

> The evidence that is accumulating suggests that learners would do well to learn the common words of the language very thoroughly because they carry the main patterns of the language. The patterns have to be rather precisely described in order to avoid confusions, but then are capable of being rather precisely deployed.
>
> At present many learners avoid the common words as much as possible, and especially the idiomatic phrases. Instead they rely on larger, rarer and clumsier words which make their language sound stilted and awkward. This is certainly not their fault, nor is it the fault of the teachers, who can only work within the kind of language descriptions that are available.
>
> Now we can have access to much more reliable information, and learners will be able to produce with confidence much more idiomatic English, with less effort involved.

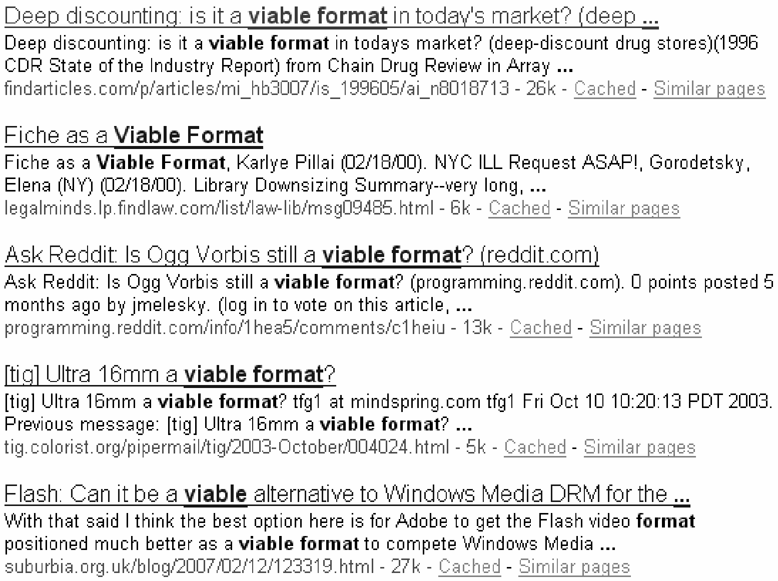
Antoinette Renouf pointed to the promising amount of possible developments stemming out of the *Cobuild Dictionary* work: (1987: 178):

> The millions of examples in the Cobuild Corpus are concrete evidence of the language, and they are available for a wide range of applications in language study, teaching and testing. The observations recorded in the database amount to a detailed study of a large and central vocabulary, and they are stored in an exceptionally flexible form.
>
> The dictionary gives rise to a number of potential spin-offs, and the lexical syllabus outlined in this chapter has a status in its own right. There is now a large collection of ideas for future publications […]

A large collection of dictionaries, lexical syllabuses (e.g. Sinclair and Renouf 1988) and Cobuild based course books (e.g. Willis and Willis 1988) has been effectively published. As it is largely known, the Cobuild project has been just one of many corpus-based research and publishing initiatives.

One should bear in mind that IBM had announced its Personal Computer/AT with a (troublesome) 20 MB hard disk drive only August 1984! CD-ROMs (available since 1984) offered a viable format for publication of corpora subsets: COBUILD went *On CD-*

ROM in 1994. In the mid-1980s, the electronic (digital) format of many of the texts which entered the corpora he was working with – as publishing was turning into a digital business – made Sinclair describe his vision of an automatically continuously fed digital corpus and of real time concordancers plugged into it, allowing for diachronic as well as for updated synchronic studies. The World Wide Web, which Berners-Lee started weaving at CERN *Conseil Européen pour la Recherche Nucléaire*) in 1991 (Berners-Lee 1999; Capra 2005: 156-164) turned that vision into reality. The web makes it possible for students and teachers around the world to easily access large corpora like, for example, the British National Corpus.[1] Yet it is the Web itself which has grown into a huge multilingual corpus, which common, largely available search tools can sieve in seconds. A simple search on Google for "viable format", for example, can return hundreds of occurrences in context, quick as a wink (Figures 1 and 2).



Deep discounting: is it a **viable format** in today's market? (deep ...
Deep discounting: is it a **viable format** in todays market? (deep-discount drug stores)(1996 CDR State of the Industry Report) from Chain Drug Review in Array **...**
findarticles.com/p/articles/mi_hb3007/is_199605/ai_n8018713 - 26k - Cached - Similar pages

Fiche as a **Viable Format**
Fiche as a **Viable Format**, Karlye Pillai (02/18/00). NYC ILL Request ASAP!, Gorodetsky, Elena (NY) (02/18/00). Library Downsizing Summary--very long, **...**
legalminds.lp.findlaw.com/list/law-lib/msg09485.html - 6k - Cached - Similar pages

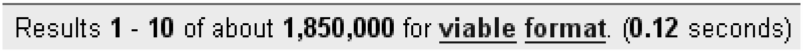
Ask Reddit: Is Ogg Vorbis still a **viable format**? (reddit.com)
Ask Reddit: Is Ogg Vorbis still a **viable format**? (programming.reddit.com). 0 points posted 5 months ago by jmelesky. (log in to vote on this article, **...**
programming.reddit.com/info/1hea5/comments/c1heiu - 13k - Cached - Similar pages

[tig] Ultra 16mm a **viable format**?
[tig] Ultra 16mm a **viable format**? tfg1 at mindspring.com tfg1 Fri Oct 10 10:20:13 PDT 2003. Previous message: [tig] Ultra 16mm a **viable format**? **...**
tig.colorist.org/pipermail/tig/2003-October/004024.html - 5k - Cached - Similar pages

Flash: Can it be a **viable** alternative to Windows Media DRM for the **...**
With that said I think the best option here is for Adobe to get the Flash video **format** positioned much better as a **viable format** to compete Windows Media **...**
suburbia.org.uk/blog/2007/02/12/123319.html - 27k - Cached - Similar pages

Figure 1. Links to the contextual occurrences of "viable format" resulting from a search on Google.

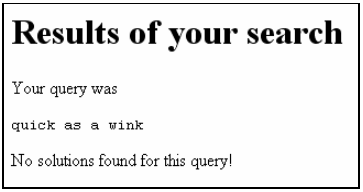
[1] Although with limited scope and range (only a random choice of up to 50 occurrences can de displayed, with no format option), the BNC can be freely accessed on line: http://www.natcorp.ox.ac.uk/.



Results **1** - **10** of about **1,850,000** for **viable format**. (**0.12** seconds)

Figure 2. Numbers and speed of the search on Google.

Let it be noted that while a search for ‘quick as a wink’ on the web version of the British National Corpus returns zero occurrences, the same query on Google results in almost 2.5 million! (Figures 3 and 4).



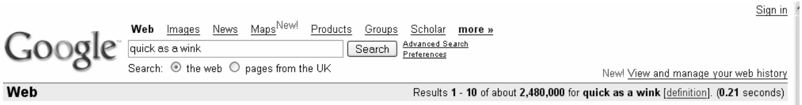
Results of your search

Your query was

quick as a wink

No solutions found for this query!

Figure 3. The Web BNC gives no occurrences for this search.


Sign in
Web Images News Maps New! Products Groups Scholar more »
quick as a wink
Search
Advanced Search
Preferences
Search: ⦿ the web ○ pages from the UK
New! View and manage your web history
Web

Results **1** - **10** of about **2,480,000** for **quick as a wink** [definition]. (**0.21** seconds)

Figure 4. The same search gives quite different results on Google.

There is of course no control on the balance of text types which make up the peculiar corpus searched by the Google engine, yet it could be argued that the size itself of the Web and the diffused origin of its very large contributing ‘constituency’ make for a much more authentically balanced mix of text types than most ‘doctored’ and smaller academically assembled corpora.

Zipf’s law on word frequency (Zipf 1949: 32-33) – which could be rephrased into the motto ‘very few words are very frequent, most words are most infrequent’– asserts its strong rule on the statistical economics of corpora, requiring a huge data base to encounter even single occurrences of any word but the few very frequent ones. Johns was among the first scholars to suggest that, for classroom use, the size rule on corpora could be bypassed by narrowing the corpus within the boundaries of specific language or ESP (Johns 1991).

A peculiar instance of restricted specific corpora is offered by learner corpora (Granger, Dagneux and Meunier 2002; Prat Zagrebelsky 2004; Martelli 2007), a fundamental contribution to the study and understanding of the interlanguage of learners – "comparative interlanguage analysis" in Sylviane Granger's definition (1998) – offering not only authentic examples of deviation from native use – 'misuse' – but quantitatively weighing 'over-' and 'under-use'.

Learner and teacher corpora have also been the basis for genre/purpose specific language teaching and course writing, such as the excellent Warwick University *EASE* (Essential Academic Skills in English) CD-ROMs (Nesi 2001). *EASE* is a series of interactive corpus-based academic English CD-ROMs specifically designed for students studying English as a second language for academic purposes in an English speaking country and focused on the skills of listening to lectures, making presentations and taking part to seminars and discussions.

The volume edited by Maria Teresa Prat Zagrebelsy (1998) on lexis and language learning appropriately devoted a section to corpora and concordances (Brodine 1998) which promised to be the most rapidly expanding area in language teaching and learning. A promise kept at university level, most of all in the research department, but in the linguistics and partially in the language classes as well. On the other hand, after a promising start, the stream of ideas and materials coming from corpus linguistics seems to have drifted out of the language classrooms in schools.

The attention to 'chunks' and 'collocations' at the basis of the lexical approach (Lewis 1993, 1997, 2000) are characteristic of a corpus-based perspective. Yet the switch, in Michael Lewis's words, from lexicalised grammar to grammaticalised lexis does not seem to have fully conquered the practice of school language teaching.

2. Learnable?

Long and Crookes (1992: 27) observed that

> Three new, task-based syllabus types appeared in the 1980s: (a) the procedural syllabus, (b) the process syllabus, and (c) the task syllabus. They are distinguishable from most earlier syllabus types by the fact that part of their rationale derives from what is known

> about human learning in general and/or second language learning in particular rather than, as is the case with lexical, structural, notional, functional, and relational syllabuses, primarily from an analysis of language or language use. In addition, while differing from one another in important ways, all three reject linguistic elements (such as word, structure, notion, or function) as the unit of analysis and opt instead for some conception of task.

Building on Wilkins's distinction between 'synthetic' and 'analytic' syllabuses(1974, 1976), they try updating his definitions (1992: 29):

> analytic syllabuses are those which present the target language whole chunks at a time, without linguistic interference or control. They rely on (a) the learners' assumed ability to perceive regularities in the input and to induce rules (or to form new neural networks underlying what looks like rule-governed behavior), and/or (b) the continued availability to learners of innate knowledge of linguistic universals and the ways language can vary, knowledge which can be reactivated by exposure to natural samples of the L2. Procedural, process, and task syllabuses are all examples of the analytic syllabus type. Wilkins (1976) classifies situational, notional, and functional syllabuses as analytic. Notions and functions are clearly linguistic units, however, isolation of which in practice always results in a synthetic syllabus, such that exercises practising requests or apologies replace exercises on relative clauses or the present perfect.

Another categorization for syllabus design they recall is White's (1988) distinction between Type A and Type B syllabuses (1992: 29):

> *Type A* syllabuses focus on *what* is to be learned: the L2. They are interventionist. Someone preselects and predigests the language to be taught, dividing it up into small pieces, and determining learning objectives in advance of any consideration of who the learners may be or of how languages are learned. […]
>
> *Type B* syllabuses, on the other hand, focus on *how* the language is to be learned. They are noninterventionist. They involve no artificial preselection or arrangement of items and allow objectives to be determined by a process of negotiation between teacher and learners after they meet, as a course evolves. […]

Once declared their preference for analytic Type B syllabuses, Long and Crookes concede that "[i]f any targetlike linguistic items are learnable separately and completely at one time, words or collocations may be the most likely candidates" yet they soon object that "[w]here syllabus design is concerned, however, problems of authenticity and learnability once again limit the potential of this effort" (1992: 32). Dave Willis (1993), co-author of the first COBUILD based course book (Willis and Willis 1988), firmly and appropriately rejects Long and Crookes' accusation of exposing the learner to no authentic samples of the target language. Yet the problem of learnability should not be easily dismissed (notwithstanding Willis's rejection of this second accusation).

Without entering the complex discussion of the mathematical theory of language learnability (Pinker 1991, 1996), a large number of practising language teachers, through their experience and observations, will support the opinion that an accurate, effective, authentic and updated 'description' of a language is no guarantee of effective learning/acquisition of the 'competence' and 'ability' allowing to communicatively 'use' that language. This seems particularly true of collocations and of the 'grammaticalised lexis' extracted from corpora. It is the result of a long tradition: for decades, if not centuries, grammar has been considered the engine of language, and lexis just the fuel. Cohorts of students – and, most relevantly, of student teachers – have been raised to eavesdrop for the clockwork ticking of such a machine in the sound of language. Moreover, the rules governing language use have been presented in a deterministic way, within a strict cause-effect correlation confirmed by a narrow parallax reading error of exceptions. Corpus linguistics has disclosed a probabilistic understanding of language events, where 'real' collocations take specific, quantistic places discretely scattered along what, from a grammatical distance, might seem like a continuum. What can be saluted as a definitive progress among the researchers in the linguistic department, leaves the alerted language teachers in the school with a feeling of frustration, since their and their students' teaching and learning experiences defeat the purpose of really 'acquiring', making their own, the lesson taught by corpus linguistics.

3. Conclusion: the quantum leap

Given that I agree with Primo Levi (1988: 77) that "you have to be careful with similes; because they may be poetic, but they don't prove much, so you have to watch your step in drawing educational or edifying lessons from them”,[2] the difficulty that the language teacher faces in lining up a sequence of classroom activities and experiences leading to a real and concrete ‘cognitive manipulation’ of the knowledge which can be distilled from a corpus – such difficulty seems very similar to the one facing physics teachers when they try to get their students, groping through a traditionally mechanical world, to grasp quantum physics.

Rinaudo (2007) lists the main difficulties that (Italian) high school students face when studying quantum physics:

- the stupor caused by conceptual complexity;
- the problem of the passage from continuum to discrete mechanics;
- the lack of macroscopic references for a visualisation of phenomena;
- the complexity of formal definition.

The list could very easily be adapted to describe the difficulties that Italian high school students face with corpus linguistics, probably just rephrasing the second line in the list as:

- the problem of the passage from (open choice) grammar to collocational lexical description and reference for language use.

Rinaudo formulates a proposal for a teaching strategy which could be summarised in two main concerns: a) tracing what learning experiences prior to the study of quantum mechanics could support rather than hinder it; b) pointing to a seminal but troublesome concept that she identifies in the ‘quantum of action’. Knocking on

[2] Levi adds: “Should the educator take as his model the smith, who roughly pounds the iron and gives it shape and nobility, or the vintner, who achieves the same result with wine, separating himself from it and shutting it up in the darkness of a cellar? Is it better for the mother to imitate the pelican, who plucks out her feathers, stripping herself, to make the nest for her little ones soft, or the bear, who urges her cubs to climb to the top of the fir tree and then abandons them up there, going off without a backward glance? Is quenching a better didactic system than the tempering that follows it? Beware of analogies: for millennia they corrupted medicine, and it may be their fault that today's pedagogical systems are so numerous, and after three thousand years of argument we still don't actually know which is best”.

the wood of Primo Levi's warning, one could similarly advocate a) a less prominent role for 'grammar rules' in language teaching, with their replacement – AMAP ASAP (as much as possible, as soon as possible) – with collocational examples from authentic corpora; b) the creation of viable corpus-based tasks for the classroom, appropriate for the students in their actual reality, drawing the focus of attention from 'what' should be learnt – substituting lexis and collocations from corpora to traditional 'grammar rules' is not sufficient – to the 'process' leading to learning and acquisition.

Such a perspective might simply be translated into the necessity to bring into the classroom not only the knowledge of the good linguist but the expertise of the competent language teaching – or, rather, language learning – methodologist. In the end, what matters in the language classroom is not how language is described as much as what students do to make it their own.

References

Berners-Lee, T. (1999) *Weaving the Web. The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*, Harper, San Francisco.

Brodine, R. (1998) "Concordanze e consapevolezza lessicale. L'uso dei corpora per l'esplorazione del lessico", in M.T. Prat Zagrebelsky (ed.), *Lessico e apprendimento linguistico*, La Nuova Italia, Firenze, pp. 171-192.

Capra, U. (2005) *Tecnologie per l'apprendimento linguistico*, Carocci, Roma.

Johns, T.F. (1991) "Should you be persuaded?", in T. Johns and P. King (eds), "Classroom concordancing", *Birmingham University English Language Research Journal* 4 [Theme issue], pp. 1-17.

Granger, S. (1998) *Learner English on Computer*, Longman, London/New York.

Granger S., E. Dagneux and F. Meunier (eds) (2002) *International Corpus of Learner English*, UCL Presses Universitaires de Louvain, Louvain.

Levi, P. (1988) *The Wrench*, (translation by W. Weaver), Abacus, London [orig. title *La chiave a stella*, Einaudi, Torino 1978]

Lewis, M. (1993) *The Lexical Approach: The State of ELT and a Way Forward*, Language Teaching Publications, Hove.

Lewis, M. (1997) *Implementing the Lexical Approach: Putting Theory into Practice*, Language Teaching Publications, Hove.

Lewis, M. (ed.) (2000) *Teaching Collocation: Further Developments in the Lexical Approach*, Language Teaching Publications, Hove.

Long M.H. and G. Crookes (1992) "Three approaches to task-based syllabus design", *TESOL Quarterly* 26 (1), pp. 27-56.

Martelli, A. (2007) *Lexical Collocations in Learner English: A Corpus-based Approach*, Edizioni dell'Orso, Alessandria.

Nesi, H. (2001) "EASE: a multimedia materials development project", in K. Cameron (ed.), *CALL: The Challenge of Change*, Elm Bank Publications, Exeter.

Pinker, S. (1991) *Learnability and Cognition: The Acquisition of Argument Structure (Learning, Development, and Conceptual Change)*, The MIT Press, Cambridge (USA).

Pinker, S. (1996) *Language Learnability*, Harvard University Press, Harvard.

Prat Zagrebelsky, M.T. (ed.) (1998) *Lessico e apprendimento linguistico*, La Nuova Italia, Firenze.

Prat Zagrebelsky M.T. (ed.) (2004) *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage,* Edizioni dell'Orso, Alessandria.

Renouf, A. (1987) "Moving on", in J. Sinclair (ed.), *Looking Up. An Account of the COBUILD Project in Lexical Computing*, Collins ELT, London/Glasgow, pp. 167-178.

Rinaudo, G. (2007) "Il mondo dei quanti. Perché introdurre la meccanica quantistica nella scuola secondaria superiore", in *Treccani Scuola*, 03.04.2007. http://www.treccani.it/site/Scuola/nellascuola/area_fisica/quanti/rinaudo.htm

Sinclair, J. (1987) "The nature of evidence", in J. Sinclair (ed.), *Looking Up. An Account of the COBUILD Project in Lexical Computing*, Collins ELT, London/Glasgow, pp. 150-159.

Sinclair, J. and A. Renouf (1988) "A lexical syllabus for language learning", in R. Carter and M. McCarthy (eds), *Vocabulary and Language Teaching*, Longman, New York/London, pp. 140-158.

White, R.V. (1988) *The ELT Curriculum. Design, Innovation and Management,* Basil Blackwell, Oxford.

Wilkins, D.A. (1974) "Notional syllabuses and the concept of a minimum adequate grammar", in S.P. Corder and E. Roulet (eds), *Linguistic Insights in Applied Linguistics*, AIMAV and Didier, Brussels and Paris, pp. 119-128.

Wilkins, D.A. (1976) *Notional Syllabuses*, Oxford University Press, Oxford.

Willis, D. and Willis J. (1988) *Collins COBUILD English Course*. Collins, London.

Willis, D. (1993) "Comments on Michael H. Long and Graham Crookes's 'Three approaches to task-based syllabus design'. A reader reacts…", *TESOL Quarterly*, 27 (4), pp. 726-729.

Zipf, G.K. (1949) *Human Behaviour and the Principle of Least Effort. An Introduction to Human Ecology*, Hafner Publishing Conpany, New York [*1972 facsimile edition*].

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 365-378 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Error analysis and learner corpora: a study of errors in the written production by English students of Italian

Aurelia Martelli – University of Turin

1. Introduction

This paper describes a study of learner errors carried out on a corpus of written texts by English students of Italian. The principles of Error Analysis (EA) and the tools and methodology of Computer-based Learner Language Research are combined in order to account for how EA can contribute to the exploitation of a learner corpus by providing a framework for the development of an error tagging system. The aim is to find out whether tagging a learner corpus for errors can be an efficient way to study characteristics of interlanguage, by helping identify and evaluate patterns in the errors of learner production.

2. Error tagging

One of the most controversial and problematic aspects of error tagging is that, unlike some other types of annotation, there is still no way of automating the tagging process. For this reason, error tagging is usually more time consuming than other types of tagging and inevitably limited to corpora of small size. Error tagging is also a type of tagging that is highly subjective, in spite of the many efforts made to ensure the highest possible level of consistency. In fact, error categories devised by different researchers are never the same and they usually reflect the type of errors that researchers are either expecting or concentrating on.[1]

For the present investigation I decided to create my own error taxonomy and tagset, even if I was aware of other taxonomies used in previous studies (see, among others, Granger *et al.* 1998; Tono 2002, 2003). There are two main reasons for this choice. First of all, many of the already existing taxonomies were designed for corpora of learner English. In working with Italian I felt it necessary to develop an *ad hoc* tagging system, appropriate for dealing with the description of errors in learner Italian. Secondly, many of the existing taxonomies have been designed for the analysis of errors by upper intermediate and advanced students while the students involved in this study were all classified as lower intermediate. Errors by students of different proficiency level are likely to differ both quantitatively and qualitatively. A taxonomy designed to account for errors by advanced students would probably have been inappropriate for tagging a corpus of lower intermediate students. In the next sections I will provide a detailed description of the procedure of collecting, tagging and exploiting the corpus.

3. Data collection

Throughout the 1998-1999 academic year at Lancaster University I collected and error-tagged a 31,000 word corpus of learner Italian (henceforth CLEI) consisting of essays produced by twenty-four students.[2]

Six group of essays were collected,[3] the first three written in the first term of the academic year (Michaelmas term) between October and mid-December, the second three produced in the second term

[1] For studies on error analysis and annotation which cover issues such as reliability of tagging schemes, automation in error tagging, development of error tagsets/editors and other methodological issues involved in error annotation see Granger *et al.* (1994); Milton and Chowdhury (1994); Dagneaux *et al.* (1996, 1998); James (1998); Yang (1998); Granger (1999, 2003); Tono (2002, 2003).

[2] All students were English mother tongue speakers, about 20 years of age and in the second year of modern European languages at the University of Lancaster. All information regarding students and the learning situation were recorded through the compilation of personal questionnaires.

[3] The essays used are between 250 and 350 words and constituted assessed coursework.

(Lent term) between January and mid-March. Essays were written at about three weeks' distance from each other. The aim was that of tracing significant changes in the writing of this group of students in the course of the academic year.

Topics of the essays vary from the description of students' families and hobbies to the discussion of stereotypes between the Italian and British people. Although the topics are quite different from each other, they are similar in that they are descriptive in nature and non-technical.

4. Error classification and taxonomy

When the tagging began the error taxonomy had not been fully developed yet. At this stage the objective was to create a system of codes (tagset) that were first of all transparent, easy to memorise and flexible enough to be modified as new types of errors were encountered. In order to satisfy these needs I opted for a hierarchical system, i.e. one based on major category codes followed by a series of more specific sub-codes. The broad categories adopted indicate the level of language involved (morphology, syntax etc.) and the part of speech. The sub-codes which follow them are more specific tags that account for alterations of the surface structure caused by word order, omissions, additions and the like.

Error categories

The six major categories correspond to six different levels of language: vocabulary (Vo), form (Fo), morphology (Mo), lexico-grammar (Xg), syntax (Sx), discourse (Ds). At the second level of classification, the first sub-code which follows the main category indicates the part of speech involved. The parts of speech indicated and their respective codes are: noun (No), pronoun (Pn), verb (Vr), adjective (Aj), adverb (Av), preposition (Pr), conjunction (Co), article (Ar), particle (Pt), compound preposition (Pa), participle (Pl), phrase (Ph).[4] From the third level of classification onwards the subcategories become more specific and change according to the type of error.

[4] The code Ph, although not referring to a part of speech, is used when an error is not limited to a single word, but when it involves a multiword expression.

Error description

Vocabulary errors include all errors that involve the incorrect choice of a word or phrase. Both lexical and functional words are taken into account. They may be caused by the choice of an incorrect word (Ch) or by the choice of a non-existent word (Ix). In both cases there can be a further level of classification which specifies whether the expression used is a false friend (Ff). Following are two examples:[5]

(1) **fare** un omicidio [VoVrCh]

(2) in città ci sono molte **facilità** [VoNoChFf]

The category of formal errors includes all errors related to orthography. They can be general spelling errors (Sp), errors regarding the (mis)use of double consonants (Db), the (mis)use of stress mark (St), the (mis)use of apostrophe (As). With the exception of the category of spelling errors (Sp) the other three sub-categories can be further divided into omission (Om) and addition (Ad) errors. For example the word *PERCHE* would be classified as [FoCoStOm] as it represents a formal error (Fo) involving a conjunction (Co) and is caused by the omission (Om) of the stress mark on the final syllable (St).

Morphological errors relate to the derivational and inflectional formation of words. Considering that Italian is morphologically a very rich language, this category was bound to be quite complex and articulated. There are different types of morphological errors that are specified at the third level of classification. The code (Dv) identifies both derivational and inflectional errors, i.e. errors in the construction of new words as well as errors in inflecting a word in order to express a grammatical contrast, as in the following example:

(3) poi le **paragonarò** col mio punto di vista [MoVrDv]

In this case the thematic vowel *a* should have been changed to *e*, the correct form being *paragonerò*.

Another more specific category of morphological error indicates whether the error involves number and/or gender agreement (Nm or

[5] All the examples presented in this and in the next sections are taken from the CLEI corpus.

Gd) as in the expression:

(4) la campagna è **tranquillo** [MoAjGd]

As far as nouns are concerned a special category had to be created to include all those errors caused not by gender agreement but by gender choice (Gc), that is to say, when the student attributed the wrong gender to a noun, as in:

(5) **la** clima **brutta** [MoNoGc]

Finally one sub-category of morphological errors referring to person agreement (Pe) had to be designed exclusively for verbs. Errors in the person agreement of verbs are rare, but nonetheless some instances were found and needed to be accounted for. An example would be the expression

(6) Io e Marku **sono** sempre usciti [MoVrPe]

The category of lexico-grammatical errors includes two different types of errors: errors in the choice of the auxiliary for compound verb forms (Ax) and errors in word order. The latter can be further divided into general position errors (Ps), which involve the incorrect positioning of words or the omission and addition of items, and colligation errors (Cg), which are caused by the violation of grammatical relations in which some sort of dependency is involved, more specifically the combination of a 'dominant' word, typically a noun or a verb, and a grammatical word, usually a preposition. In the latter case specific codes indicate whether the dependency relation which is violated occurs between a noun and its preposition (Np), an adjective and its preposition (Ap), a verb and its preposition (Vp) or an infinitive and its preposition (Ip). At the final level of specification the codes Om, Ad or Ch are added to specify whether the error involves the omission, addition or choice of the preposition in question. Following are some examples:

(7) quando penso **degli** Italiani [XgPrCgVpCh]

(8) allora **provo andare** a casa mia [XgPrCgVrOm]

(9) non è molto facile **a** leggere [XgPrCgApCh]

The category of syntax in this study is limited to verbal syntax. Syntax errors concern the choice between finite and non-finite forms (Fn), active and passive forms (Vc), choice of mood (Md) and choice of tense (Ts). Following are some examples:

(10) Si potrebbe evadere dalla città e **va** in campagna [SxVrFn]

(11) Ho bisogno di **circondare** da amici [SxVrVc]

(12) Se **abito** a Londra prenderei la metropolitana [SxVrMd]

(13) Nel 1997 **venivo** a Lancaster [SxVrTs]

Finally a category of 'discourse' errors was created for all those expressions which could not be interpreted and for which the intended meaning could not be reconstructed, as the following example shows:

(14) Un vantaggio importante sarei molto cerca a molti posti che sono necessario a vivere.

The following table summarises the categories.

Main Category	Sub-category 1	Sub-category 2	Sub-category 3	Sub-category 4
Vo	Part of Speech	Ix Ch	Ff	
Fo	Part of Speech	Sp Db St As	Om/Ad	
Mo	Part of Speech	Dv Gd Nm Pe		
Xg	Part of Speech	Ax Ps Cg	Om/Ad Np/Ap/Vp/Ip	Om/Ad/Ch
Sx	Vr	Fn Md Ts Vc		
Ds	Ph			

Table 1. The error taxonomy.

5. Data exploitation

Errors were classified by attaching a tag contained in square brackets before the word or expression considered erroneous. In order to analyse the tags WordSmith Tools was used, more specifically the Wordlist tool to find the total number of errors in the corpus and the Concord tool to analyse more closely the error types which revealed some significant or unusual pattern.

Statistics

The first step in the analysis of error frequency was to break down the errors into six groups corresponding to the six major error categories in the taxonomy. The following table summarises the results. The names Mich 1,2,3 and Lent 1,2,3 correspond to the different Word files in which the corpus was stored. Each file corresponds to a group of essays and the name of the file reflects the chronological order in the essays were produced:

	Sx	**Mo**	**Xg**	**Vo**	**Fo**	**Ds**	**Total errors**
Whole corpus	232	997	800	935	407	12	3,383
Mich1	21	25	39	45	45	0	175
Mich2	34	82	81	90	49	2	338
Mich3	16	81	58	96	31	1	283
Lent1	48	174	169	228	87	1	707
Lent2	63	230	193	225	87	6	804
Lent3	50	405	260	251	108	2	1,076

Table 2. Distribution of errors in the CLEI corpus.

Table 2 provides a general idea of the distribution of the different types of errors in the corpus. It may, however, be misleading, as it does not account for the number of words that each file contains. In order to provide a more reliable picture of the actual proportion of errors contained in each essay group the percentage of errors per 100 words was calculated.

File name	Words	Errors	Percentage of errors
Mich1	1,505	175	11.6%
Mich2	3,967	338	8.5%
Mich3	2,874	283	9.8%
Lent1	5,923	707	11.9%
Lent2	7,672	804	10.5%
Lent3	9,351	1,076	11.5%
Whole corpus (CLEI)	31,292	3,383	10.8%

Table 3. Percentage of errors in each essay group.

It seems that, in terms of the quantity of errors, there is no progress in student production. The percentage of errors does not decrease as one might expect, or at least hope, considering that essays were collected at some weeks' distance from each other. This fact does not of course imply that students did not make any kind of progress. One of the factors that Table 3 does not account for, and that might explain the invariability in the proportion of errors, is that students are taught more as they progress and therefore that they have more linguistic material to cope with.

This is a possibility that was taken into consideration by analysing the significance of error categories in terms of the frequency of error for each type. In other words, I wanted to see whether one or more of the categories could be described as characteristic of one particular group of essays, and therefore be associated with a specific stage in the learning process.

In order to do this a chi-squared test was performed. The chi-squared technique provides a "test of the significance of the difference in proportions […] it can be thought as a test of association between the categories used" (Robson 1973: 85). In the case of the present study, it can be used to verify whether it is possible and plausible to associate the variable of time, i.e. the chronological sequence in which the essays were written, with the number of errors of each type.

The first step consisted in calculating the expected frequency of the different error categories.[6]

[6] The expected frequency (*E)* is calculated by multiplying the total number of errors in a single group of essays by the total number of a specific type of error in the entire corpus and then dividing the result by the total number of all errors in the corpus.

	Sx	Mo	Xg	Vo	Fo	Ds
Mich1	12.00	51.57	41.38	48.37	21.05	0.62
Mich2	23.18	99.61	79.93	93.42	40.66	1.20
Mich3	19.41	83.40	66.92	78.22	34.05	1.00
Lent1	48.48	208.36	167.19	195.40	85.06	2.51
Lent2	55.14	236.94	190.13	222.21	96.73	2.85
Lent3	73.79	317.11	254.45	297.39	129.46	3.87

Table 4. Expected frequencies of the error categories.

In comparing the expected frequency and the actual data, or the observed frequency (O), it becomes evident that in most cases there is a more or less noticeable difference between the two. The discrepancy between the expected frequency and the observed frequency is then used to compute the chi-squared statistic (X) by using the following formula:

$$X^2 = \Sigma \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}$$

	Sx	Mo	Xg	Vo	Fo	Ds
Mich1	6.748	13.693	0.137	0.234	27.236	0.621
Mich2	5.052	3.114	0.014	0.125	1.7089	0.535
Mich3	0.598	0.069	1.190	4.044	0.273	0.000
Lent1	0.005	5.666	0.021	5.438	0.044	0.907
Lent2	1.121	0.204	0.043	0.035	0.978	3.475
Lent3	7.667	24.362	0.121	7.236	3.555	0.865
Observed chi-squared	127.134					

Table 5. Observed chi-squared.

In order to assess the resulting chi-squared value for significance it is necessary to compare the value obtained with the chi-squared reference table reported in most textbooks dealing with statistical operations. For the present study the reference values are those reported in Sholefield (1995) which provides reference chi-squared values at the 5% significance level.[7] The chi-squared value obtained and reported in Table 5 (127.134) abundantly exceeds the reference value found in Sholefield (37.653). Any time that the obtained chi-squared value exceeds the chi-squared value in the reference table, it can be said that there is association between the categories used, in this case the type of error and the time of production. Thus it can be said that the time at which the essays were written affects the nature of the errors, that it to say, different types of errors characterise different groups of essays written at different times in the academic year.

The two error categories in which the discrepancy between the expected and the observed frequency is more evident are those of syntax and morphological errors. In the following sections syntax errors in the Mich1, Mich2 and Lent3 files will be taken into consideration, since syntax is the category of errors in which the discrepancy between observed and expected frequency was assigned a higher and more significant value.

6. Syntax errors

In order to carry out a close analysis of the syntax errors in the above-mentioned files the Concord tool was used. By using the wild card option it was possible to retrieve all occurrences of syntax errors in the individual files. The next step consisted in subdividing all syntax errors according to the subcategories that indicate the different type of error involved. This part of the analysis had to be done manually by reading through all the concordances and counting the occurrences of the different error types. Table 6 illustrates the results:

	Md (mood)	**Ts (tense)**	**Vc (voice)**	**Fn (finiteness)**	**Total errors**
Mich1	2	17	2	0	21
Mich2	9	22	0	3	34
Lent3	28	11	0	11	50

Table 6. Subdivision of syntax errors.

[7] A 5% significance level means that the chi-squared test performed is reliable in 95% of cases.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

The chosen sub-categories are those of tense and mood.

Tense errors

Tense errors are significantly frequent in the first two essay groups where they constitute approximately 81% and 65% of all syntax errors. They are less frequent in the last essay group where they constitute only 22% of all syntax errors. A possible explanation is that tense errors are more frequent earlier in the year when students are still attempting to master some tense distinctions, while they are more or less overcome by the time the last essay is written, at the end of the Lent term.

In the Mich1 file thirteen out of the seventeen tense errors (76%) are caused by inappropriate use of the *imperfetto* (imperfect) and *passato prossimo* (present perfect) tenses. The same can be said for nine out of the twenty-two tense errors (41%) found in Mich2. In such cases students choose the imperfect instead of the present perfect tense, or vice versa, as in the following example:

(15) Quando **siamo stati** giovani abbiamo abitato…

In Italian the two tenses mentioned above have specific uses and express different things about the past. Generally speaking, it can be said that the *passato prossimo* is a narrative tense and refers to specific actions completed in the past, independently of whether the action lasted a long or short time, or whether it took place once or a number of times. The *imperfetto*, on the contrary, is a descriptive tense and refers to how people and situations were in the past (mostly in terms of physical, mental, emotional conditions and states of being) and to actions which were habitual or in progress at a specific time, but does not refer to an action being completed.

Such distinctions are hard to master for English students basically because the tense system in English is structured differently. In English, tense contrast is between the simple past and present perfect. The question is not so much whether a past action is portrayed as completed or not, or whether it refers to a state of being, but rather to whether there is any relationship between the past event and the present time. In expressing past events and states, English students cannot rely on the tense structure of their mother tongue. Thus, it seems that the difficulties in dealing with specific structures can be ascribed to fundamental differences in the way certain distinctions are conceptualised and therefore grammaticalised in the mother tongue and in the target language.

Tense errors in the last set of essays, with the exception of one error, do not involve the above-described distinction. Instead they are caused by failure to master the rules regarding the 'tense sequence', as in the following example:

(16) Comunque è necessario che gli stereotipi **fossero** infranti

It is not surprising that this type of tense error does not occur in the first two groups of essays since the issue of tense sequence in Italian is related to the use of the subjunctive and the formation of hypothetical if-clauses. Both topics are not dealt with until the Lent term, so that they should not cause problems until then.

Mood

Beside its strict relationship with tense errors, the issue of the subjunctive mood is in itself a problematic one for English students. In the final group of essays, written shortly after the subjunctive mood had been covered in the classroom, mood errors constitute 56% of syntax errors, and of these 86% are related to the use of the subjunctive (24 out of the total 28 mood errors). Errors are made because students often use the subjunctive when the indicative is required, or because they use the indicative mood when the subjunctive is required, as in:

(17) Voglio segnalare che gli stereotipi **siano** interessanti

(18) Penso che **è** perché loro hanno un'alimentazione sana

Another interesting observation can be made about the use of the subjunctive in the first two essay groups. As mentioned above, the subjunctive is not dealt with until the second term of the academic year. In spite of this fact, there is one error involving the subjunctive in the first essay group and eight in the second essay group. It would be interesting to investigate the possible reasons why students use, or try to use, a structure which they have not been taught. What is also strange is that students using the subjunctive in the first two essay groups make mistakes as far as its use is concerned, just as in the final group of essays, but conjugate the verb correctly. This phenomenon would be worthwhile investigating further. It could be done, for example, by asking the students themselves why they used the subjunctive and what type of meaning they were trying to convey, or by analysing how the various language materials that students come into contact with outside the classroom (videos, magazines, conversations with Italian native speakers) might encourage them to pick up language structures that they have not yet encountered in the learning environment.

7. Conclusion

Through the analysis presented above I tried to trace changes in the type and number of errors which occurred as students progressed through the academic year. Statistical tests for significance have proved a useful tool for the identification of the most ‘significant’ areas, the ones worth looking at more closely. Overall it can be said that applying error analysis to a corpus was useful for both testing some already existing hypotheses and intuitions, and for revealing aspects of language use that had not been considered. Certainly there are many ways in which the present analysis could have been improved. First of all, a larger corpus would have provided the basis for more reliable analysis of the data. Secondly, had the data been not only error annotated, but also part of speech tagged, a greater number of searches could have been carried out such as verifying how many inflectional errors had been made in respect to, for example, the total number of nouns present in the entire corpus. Finally, had there been more than one analyst working on the project, the tagging process could have been crosschecked to ensure consistency and reliability.

In spite of the above mentioned limitations, the present study has nonetheless revealed that two different ways of approaching learner language, error analysis and computer-based learner language research, can indeed be combined to complement each other.

References

Dagneaux, E., S. Denness, S. Granger and F. Meunier (1996) *Error Tagging Manual Version 1.1*, Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve.

Dagneaux, E., S. Denness and S. Granger (1998) "Computer-aided error analysis", *System: An International Journal of Educational Technology and Applied Linguistics* 26 (2), pp. 163-174.

Granger, S. (1999) "Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus", in H. Hasselgård and S. Oksefjell (eds), *Out of Corpora. Studies in Honour of Stig Johansson,* Rodopi, Amsterdam, pp. 191-202.

Granger, S. (2003) "Error-tagged learner corpora and CALL: a promising synergy", *CALICO* 20 (3), pp. 465-480.

Granger, S. and F. Meunier (1994) "Towards a grammar checker for learners of English", in U. Fries and G. Tottie (eds), *Creating and Using English Language Corpora*, Rodopi, Amsterdam, pp.79-89.

James, C. (1998) *Errors in Language Learning and Use: Exploring Error Analysis*, Longman, Essex.

Milton, J. and N. Chowdhury (1994) "Tagging the interlanguage of Chinese learners of English", in L. Flowerdew and A.K. Tong (eds), *Entering Text*, The Hong Kong University of Science and Technology, Hong Kong, pp. 127-143.

Robson, C. (1973) *Experiment, Design and Statistics*, Penguin, Harmondsworth.

Scholfield, P. (1995) *Quantifying Language*, Multilingual Matters, Clavedon.

Tono, Y. (2002) *The Role of Learner Corpora in SLA Research and Foreign Language Teaching: The Multiple Comparison Approach*, Unpublished Ph.D. Thesis, Lancaster University, UK.

Tono, Y. (2003) "Learner corpora: design, development and applications", in D. Archer, P. Rayson, A. Wilson and T. McEnery (eds), *Proceedings of the Corpus Linguistics 2003 Conference*, UCREL, Lancaster University, pp. 800-809.

Yang, D. (1998) "A quantitative approach to the IL errors in Chinese EFL learners' written production", in S. Granger and J. Hung (eds), *Proceedings of the First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, The Chinese University of Hong Kong, pp. 122-124.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*
©2008 Polimetrica International Scientific Publisher Monza/Italy

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 379-397 © 2008 Polimetrica International Scientific Publisher Monza/Italy

The Role-Play Learner Corpus: a resource for investigating learner language

Maria Cecilia Rizzardi, Luciana Pedrazzini and Andrea Nava*
University of Milan

1. Introduction

With the advent of computer learner corpora that document learner language, a new and exciting field of research has opened. Both applied linguists and teachers of foreign/second languages are now in a position to "use the methods and tools of corpus linguistics to gain better insights into authentic learner language" (Granger 1994: 25). They may exploit the construction of learner corpora both to investigate learners' performance/competence and its development over time and to improve foreign language teaching and testing techniques.

It is a fact that learner corpus research to date has fed primarily into descriptions of learner language. However, new interesting applications of learner corpora have been explored recently both in the investigation of language learning processes and in the improvement of language teaching syllabi, materials and methodology. A short update on these recent developments is provided in the second section of this paper.

In the context of learner corpus research, the term 'learner' refers "either to somebody learning a foreign language or to a foreigner learning the language in a country where it is spoken natively"

* The article was planned jointly by the three authors. However, it must be pointed out that the introduction and section 4 were written by M.C. Rizzardi, sections 3 and 6 by L. Pedrazzini and sections 2 and 5 by A. Nava.

(Nesselhauf 2004: 128),[1] while the term ‘corpus’ refers to a collection of samples of the language used by learners to perform some kind of task. This raises the question of how to go about eliciting and collecting data and more specifically what kind of ‘tasks’ may stimulate the most natural language from learners. These issues are tackled in the third section of the paper.

Within this field of research, in 2002 a project was initiated at the Dipartimento di Scienze del Linguaggio e Letterature Straniere Comparate (Università degli Studi di Milano) to compile a small-scale corpus of spoken learner language. It started as part of an action research process: we had introduced the use of the role-play technique in the oral exams of first-year students of English and wanted to assess how effective the new approach was. The purpose of our research was both linguistic and pedagogical, as summarized in the following research questions:

- What features of learner spoken language are produced by our students through role-play tasks?
- Does the language in the corpus offer any evidence for what teachers of English generally and rather vaguely consider major and unacceptable errors?

The project is described in the fourth section of the paper, while the fifth section provides an initial analysis of the role-play data and attempts to answer the two research questions. The concluding part of the paper explores the role of learner corpus research in foreign language pedagogy and in teacher development giving examples of activities aimed at fostering EFL learners’ and trainee teachers’ language awareness.

[1] Recently compilation of non–native speaker corpora of English as a Lingua Franca (ELF) has begun with the aim of recording and analyzing communication between speakers of English from a variety of first language backgrounds. Two main corpora of this type are (a) VOICE, the Vienna-Oxford International Corpus of English, which aims to provide a general basis for analysis of ELF talk on all linguistic levels (Seidlhofer 2001, 2005) and (b) ELFA, a corpus of English as Lingua Franca in Academic Settings (Mauranen 2003).

2. Applications of learner corpora: an update

In this section we will attempt to provide an update on recent developments in the applications of learner corpora, focusing in particular on materials design and the influence of recently developed spoken corpora on SLA (Second Language Acquisition) research.

The main use to which learner corpora have hitherto been put is the description of learner language. Although over the past decade the range of phenomena investigated has widened to include aspects of discourse and pragmatics, notably with reference to spoken language (see below), the underlying approaches that have been taken to their analysis have remained those of Contrastive Interlanguage Analysis and Computer-Aided Error Analysis.[2] Both approaches view native-like competence as the target to which language learners aspire and rely on the availability of control corpora of native-speaker language and, less often, of the learners' L1s.

With regard to English control corpora, the issue has been raised of what type of English should indeed be taken as 'benchmark'. It may be justifiable to compare the output of the ICLE[3] contributors with essays written by British and American students, since ICLE has involved English language specialists (undergraduates of degree courses in English language and literature). However, given the acceptance that the existence of a variety of English as a Lingua Franca has started to gain among applied linguists, it is perhaps more controversial whether BANA[4]English corpora should be taken as the yardstick by which to describe corpora of the interlanguage of, for example, lower-level EFL secondary school students (Tan 2005).

[2] Contrastive Interlanguage Analysis and Computer-Aided Error Analysis were developed in the Centre for English Corpus Linguistics at the Catholic University of Louvain in the 1990s. The former involves two types of comparison, i.e. between native speech and non-native speech (NS *vs* NNS) and between two or more varieties of non-native speech (NNS *vs* NNS) (Granger 1998, 2002). The latter uses a specially devised type of computer corpus annotation for the analysis of learner corpora (Dagneaux *et al.* 1998).

[3] The International Corpus of Learner English (ICLE) is a computerized corpus of argumentative essays on different topics written by advanced EFL students from 19 different mother tongue backgrounds (Granger *et al.* 2002).

[4] BANA is an acronym for Britain, Australia, North America (Holliday 1994).

Alongside their use for purposes of learner language description, there has also been growing awareness over the last few years of the need to explore possible applications of learner corpora in language teaching (see Conclusion) and SLA.

The aspect of language teaching where the impact of learner corpora has been greatest is undoubtedly materials design. Early 2007 saw the publication of the second edition of a major monolingual dictionary for advanced learners of English as a foreign language[5] which contains several 'Get it Right' boxes highlighting frequent and widespread errors that advanced learners of English tend to make in their writing. This resource was developed by researchers from the Centre for English Corpus Linguistics at the Catholic University of Louvain and exploits findings from the ICLE project. Apart from lexicography, which is the first and most obvious outlet of corpus-informed language research, coursebooks too seem to have started to benefit from learner corpora – extracts from the Cambridge Learner Corpus, for example, are included in the series Objective IELTS (2006), where they are used as the springboard for error-editing activities (Black 2007). In other areas of language teaching (e.g. syllabus design, classroom methodology), however, the impact of learner corpus research has been less noticeable and there is thus still much scope for experimentation and development (see below).

Advances in the application of learner corpora to the study of language learning processes have been mostly the result of the widening of the types of corpora available. As pointed out by Myles (2005a), the fact that early learner corpora traditionally consist of written cross-sectional data, often unannotated, has conspired against their exploitation in SLA research. The last few years have, however, seen a reversal of this trend, as the number and size of learner corpora of spoken language have increased. Apart from the well-known LINDSEI[6] project (Pulcini 2004), whose size has grown considerably due to the addition to the database of new L1 subcorpora, a major new project has been initiated whose aim is to

[5] *Macmillan English Dictionary for Advanced Learners*.

[6] The Louvain International Database of Spoken English Interlanguage (LINDSEI) is a corpus of spoken learner language featuring interviews with advanced EFL learners from different language backgrounds.

investigate the interlanguage development of second language learners of French (French Language Learner Oral Corpora – FLLOC).[7] What characterises this database is the presence of a mixture of cross-sectional and longitudinal data as well as the fact that transcription and analysis have been carried out using the software developed for the CHILDES system, which has gained wide currency not only in first language acquisition, but also in language pathology and SLA research (Myles 2005a).

It goes without saying that the wider availability of spoken corpora has in the first place prompted exploration and description of the features which characterise spoken learner language, in keeping with similar research being conducted with native speaker corpora (e.g. McCarthy 1998; O'Keeffe and McCarthy 2007). Pulcini and Damascelli (2003) and Pulcini and Furiassi (2004), for example, among the earliest studies exploiting a completed LINDSEI subcorpus, have investigated pragmatic aspects of the spoken interlanguage of Italian EFL learners, focusing in particular on discourse markers.

A more wide-ranging effect of the development of spoken corpora has, however, been the recent interest in the learner corpus methodology among SLA researchers, who tended to be rather 'underwhelmed' by the potential of early learner corpora. Indeed, the FFLOC database project has already resulted in studies into the emergence of syntax in early French interlanguage (Myles 2005b; Rule and Marsden 2006), and it would not be overoptimistic to predict that more research will certainly follow thanks to the availability of the entire database on the Internet.

Alongside spoken corpora, great potential appears to be offered by multimedia learner corpora (e.g. the MAELC)[8] and corpora of computer-mediated communication (e.g. the Telekorp),[9] which

[7] The FLLOC database is made up of five oral corpora gathered in various institutional settings in the UK and continental Europe and representing instructed learners of L2 French from complete beginners to final-year university undergraduates (Myles 2005a).

[8] The Multimedia Adult ESL Learner Corpus (MAELC), compiled at Portland State University and Portland Community College, features approximately 5000 hours of video and audio recordings of classroom instruction involving over 1000 low-level ESL learners (Reder *et al.* 2003).

[9] The Telekorp (Telecollaborative Learner Corpus of English and German) project, developed at the Centre for Language Acquisition, Penn State University, is a

should enable SLA researchers to more easily track the emergence of interlanguage features in learners' output. The latter type of corpus, in particular, allows for synchronous data collection, thus obviating the need for a time-consuming transcription process. Given the advantages they offer over more traditional learner corpora, it is not too far-fetched to predict a burgeoning of research activity using these types of corpora in years to come.

3. Learner language: eliciting and collecting data

This section explores the methodologies for eliciting and collecting good quality learner corpus data provided by SLA research.

The study of learner language in a corpus provides information about second language learners' competence and how it develops over time. However, as learners' underlying linguistic competence is considered to be essentially of the implicit kind and is not open to direct inspection, researchers are forced to infer competence from some kind of performance. This raises the question of what kind of performance provides the most reliable and valid source of information, given the assumption that the validity of a data collection method is best established when data reflect as closely as possible 'natural' language use, that is the kind of use for which language is designed and acquired (Ellis and Barkhuizen 2005: 21). The notion of 'authenticity', usually viewed as an assumed characteristic for a corpus (Sinclair 1996),[10] is in fact very controversial with regard to learner language (Granger 2002). However, even if it is true that task variables are imposed on the learner, the kind of activity taking place in a teaching and learning context can be considered as 'natural' within that context and so can learners' language use.

Let us now consider the issue of learner corpus construction. The key factor is the nature of the sample itself which depends on how it is elicited and collected. Learner corpora generally rely on clinical elicitation methods, i.e. "getting the informant to produce data of any

bilingual longitudinal corpus consisting of e-mail and chat interactions between American university students of German and German EFL trainee teachers (Belz 2004; Belz and Vyatkina 2005).

[10] See Mauranen (2004: 91-94) and Tan (2005) for a summary of the main issues involved in the debate on 'corpus authenticity'.

sort" as opposed to experimental elicitation which involves "getting the informant to produce data incorporating particular features which the linguist is interested in at the moment" (Corder 1976: 69). The distinction between clinical and experimental elicitation has its counterpart in the distinction between 'task' and 'exercise' and bears some relevant implications in corpus construction and data analysis. Clinical elicitation involves the use of tasks where learners are

> primarily concerned with message conveyance, need to use their own linguistic resources to construct utterances, and are focused on achieving some linguistic outcome. (Ellis and Barkhuizen, 2005: 23)

The elicitation instrument is thus designed to provide a context for learners to speak or write in the L2 in a purposeful way and involve some kind of task.[11]

Task variables together with learner variables are important features in the design of learner corpora:

> the use of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables. (Granger 2002: 9)

For example, in the ICLE, learner variables refer to the learners' age, their learning context, their mother tongue, their L2 proficiency level, etc. while task variables take into account genre, medium, length of text and task settings, such as time limit, use of reference tools and whether the task is part of an exam (Granger 2003).

Different tasks have been used by SLA researchers to elicit samples of learner language: some involve social interaction and result in dialogic discourse, such as communication gap tasks, oral interviews and role-plays; others are performed by individual learners and provide examples of monologic discourse, such as text-reconstruction and picture composition tasks. SLA researchers have investigated how task design variables influence fluency, accuracy and complexity of performance (Bygate *et al.* 2001; Ellis 2003). For example, tasks that provide contextual support are about familiar or involving topics; they pose a single demand; they are

[11] Among recent studies on task-based learning and teaching see Skehan (1998); Bygate *et al.* (2001); Ellis (2003); Nunan (2004); Willis and Willis (2007).

closed and have a clear inherent structure of the outcome. All these conditions are likely to promote fluency in this type of tasks. On the other hand, tasks that do not provide contextual support, are open and have a clear inherent structure with opportunity for planning lead to more accurate language use (Ellis 2003: 127).

Ellis and Barkhuizen (2005: 26-30) describe the different ways of collecting data. Some of these are also used in learner corpus research. Written samples of learner language are relatively permanent and, for this reason, easier to collect. They can be organized in a set of sub-samples, such as those that make up the ICLE (Granger 2003). Oral samples can be collected through audio or video recording. In most situations, audio recording, especially using radio microphones or mini-disc recorders which learners carry in their pockets, is likely to provide the best data. Recorded data need to be transcribed according to a specific method of transcription which varies according to how 'broad' or 'narrow' it is.[12] The fact that samples are produced during an examination, such as in the ICLE, the Standard Speaking Test (SST) Corpus and the Role-Play Learner Corpus described in this paper, can be viewed as another variable that should not be overlooked in the process of data collection.

SLA research provides interesting insights into methods of collecting data, with particular attention to the problem of 'construct validity', which refers to the "extent to which the data provide information that can shed light on how learners acquire an L2" (Ellis and Barkhuizen 2005: 47). The validity of a particular data collection method can be achieved with reference to a clear and specific goal for the research. Some factors concerning data elicitation and collection may impact on the characteristics of learner language samples suggesting that learner corpus researchers should take particular care that design variables match the research questions to be pursued.

[12] A broad method simply provides a written record in standard orthography, perhaps noting major pauses, while a narrow method will indicate pause length, filled pauses, phonetic features, overlapping speech. See the system used for the LINDSEI.

4. The Role-Play Learner Corpus

Why did we start collecting a learner corpus? As stated in the introduction, we had introduced the use of the role-play technique in both our teaching and testing programme with first-year university students of English and wanted to assess the validity of this technique to test students' spoken interaction skills. We had chosen this type of task, because we thought that, even within the constraints of playing given roles, students are still free to make their own linguistic choices and to organize their discourse as they want and are able to do. Somehow, the role-play technique "requires the language learner to do what native speakers do with discourse" (Oller 1979: 186). Furthermore, we appreciated the fact that examiners are free to observe the students' performance, not being involved in the interaction themselves. At the same time we found out that, compared with standard corpora of native English, interlanguage corpora may be relatively small. This aspect of learner corpora gave us the idea that we could cope with the collection of a learner corpus ourselves.

Data for the Role-Play Learner Corpus have been collected according to a specific format, details of which are given below:

- the participants are all first-year university students of English, mainly Italian mother tongue speakers. The Role-Play Learner Corpus can thus be described as a quasi-longitudinal corpus (Granger 2002: 11) as it collects data from a homogeneous group of learners at different levels of proficiency (Common European Framework of Reference: levels B1– B2 – C1);
- the participants are told that they are taking part in a research project and that their interaction will be recorded;
- they work in pairs. They are given one role-card each. After reading it, they carry out a short conversation. The role-play usually lasts less than 5 minutes;
- the role-plays are about familiar topics, such as planning holidays among friends, according to the B1 level of the Common European Framework of Reference (Council of Europe 2001: 74);

- the examiner grades the participants' performance according to 'range', 'accuracy', 'fluency', 'interaction' and 'coherence' (Council of Europe 2001: 28-29);
- the participants fill in a learner profile questionnaire providing background data on age, sex, educational background, time spent in English speaking countries, etc;
- the data is transcribed according to the LINDSEI corpus guidelines;

So far 114 first-year undergraduate students of English have been recorded while interacting in 57 role-plays. The Role-Play Learner Corpus currently stands at approximately 28,000 words, but data assembly is still going on.

5. Analysing the role-play learner language

As data collection for the Role-Play Leaner Corpus is still in progress, findings from in-depth data analyses are not yet available. However, initial manual coding[13] of part of the data has been attempted, and two exploratory studies investigating selected aspects of the learners' spoken discourse have been initiated (Rizzardi *et al.* 2004; Nava 2005). In the following section, we will seek to provide tentative answers to our research questions on the basis of the issues that have arisen from the initial exploration of the data.

The first research question that we set out to investigate is aimed at identifying the features of spoken discourse (such as 'tails' and listener response tokens) that first-year students at our university produce in carrying out role-plays. The role-play tasks were designed in such a way as to trigger production of transactional language (speakers have to exchange information in order to reach an outcome – see above). However, it is a fact that no natural interaction is exclusively transactional and features of what O'Keeffe and McCarthy (2007) call "relational language"[14] appear

[13] Despite the advantages that automatic analysis software affords learner corpus researchers, manual coding and analysis may still be the best option if qualitative-based investigations are carried out (Granger 2002: 15).

[14] O'Keeffe and McCarthy (2007: 159) use this label to refer to all those linguistic devices (such as conversational routines, small talk, discourse markers, hedging

to crop up in any sort of oral communicative exchange. According to Carter and McCarthy (2003: 119), the appropriate display of relational language is one of the criteria which mark out the "successful user of English", whose performance is "judged (…) by how well they communicate, including how well they fit what they say to the needs of their listener(s)". It is thus of some interest to explore to what extent our participants resort to 'relational' linguistic features in carrying out the role-play tasks.

Analysis has so far concentrated on the identification of 'tails' in our corpus. These are linguistic devices which, as in the example below from the CANCODE corpus,[15] are used in native speaker interactive discourse to orient the message to the listener, in other words, they are used to clarify aspects of the message (for the listener's sake) or to create an atmosphere of informality and an affective bond between speaker and listener. As such, they are a prime example of relational language. Preliminary findings reported in Nava (2005) point to an underuse of this feature of relational grammar by our participants (see an example of a tail from the Role-Play Learner Corpus below), which was accounted for by several possible causes – the fact that Italian lacks the whole range of tail structures that English has, that features of spoken grammar are rarely (if ever) made the subject of explicit teaching and that the role-play interaction tasks on which the Role-Play Learner Corpus is based are carried out under exam conditions and it is admittedly difficult in these circumstances to expect completely natural language use.

(1) She was a character she was really …
(CANCODE)

(2) <B2> to the mountains . oh no: I don't want to go absolutely to the mountain I prefer to go to the sea . it's more relaxing <\B2>
<B1> it's boring the sea <\B1>
(Role-Play Learner Corpus)

and vague language) which "create and maintain good relations between the speaker and the hearer".

[15] The Cambridge and Nottingham Corpus of Discourse in English (CANCODE) consists of 5 million running words of informal spoken English (McCarthy 1998).

Connected to this issue of relational language is the concept of "listenership" (O'Keeffe and McCarthy 2007), which shifts the focus from how well people cope in oral interactions as speakers onto their role as listeners. Corpus-based research summarised in O'Keeffe and McCarthy (2007) shows that native speakers' spoken interactions are 'littered' with listener response tokens (minimal units such as *MMM, YEAH* and longer tokens such as *RIGHT, FAIR ENOUGH, IS THAT SO*?) whose overarching function is again pre-eminently affective.

Research also seems to indicate that response tokens tend to be underused by non-native speakers of English – for example, the minimal token *YEAH* occurs one fourth of the times in a corpus of non-native "successful users of English" (Prodromou 2003) than it does in a corpus of native speaker spoken English (the CANCODE). Preliminary investigation of the Role-Play Learner Corpus data has pointed to a dearth of listener response tokens in the output of first-year students at our university. In the following extract, for instance, some sort of affective response would have been expected from B2 after B1 has explained that she cannot afford the type of holiday proposed by B2. Instead, B2 rather abruptly asks a question about the length of the holiday:

(3) <B1> I don't know we we have to ask the a=agency how much it cost because you know I can't spend too much because . I haven't earned a lot this year <\B1>
<B2> and how many weeks we can .. <\B2>
(Role-Play Learner Corpus)

What is exemplified in this extract can also be viewed from a more general pragmatic perspective – that of Grice's Cooperative Principle. Participants in the Role-Play Learner Corpus experiment appear to unwittingly flaunt Grice's Maxims – sometimes saying too little, as in the example shown above, sometimes providing unnecessary detail. In the following extracts, for example, the speaker introduces information which the hearer should already be familiar with as they are playing the role of close friends planning a holiday together:

(4) <B1> [yes I suggest our going to Messico and Guatemala because you know we are studying Spanish and I think it's it's .. it's useful for us to to to speak Spanish with other persons <\B1>
(Role-Play Learner Corpus)

(5) <B1> no in Scotland no because you know that I don't like cold weather . I prefer hot weather .. since we live in Milan and em Milan you know it's . it's it's not a hot city and em <\B1>
(Role-Play Learner Corpus)

The second research question is an attempt to view the interlanguage samples gathered in the Role-Play Learner Corpus from the EFL teacher's point of view – do we find evidence of what teachers consider major and unacceptable errors? As pointed out above, the usefulness of comparing learner corpora with BANA English has been queried in recent years. However, given that the students whose output we have been collecting are future specialists in English language and (British/American) literature, we feel it is warranted to carry out error analyses taking as baseline standard English spoken in BANA countries.

Preliminary findings indicate that while instances of those that are usually regarded by EFL professionals as serious mistakes are widely represented in our corpus, so are errors of which EFL teachers are often less aware. Examples of the first type of error are shown below – dropping the third person present simple morpheme, placing adverbials between verb and direct object, use of the determiner *THE* for general reference, ellipsis of the subject pronoun:

(6) <B1> … she love much pets but em … <\B1>
(Role-Play Learner Corpus)

(7) <B2> … I like very much the gothic art and … <\B2>
(Role-Play Learner Corpus)

(8) <B2> … em for a person em is important to .. to change the direction of em of the life <\B2>
(Role-Play Learner Corpus)

In addition, even a superficial browse through the corpus transcripts cannot fail to reveal a large number of errors of the second type, i.e. those that an EFL teacher would probably not–include in a list of serious student mistakes, as shown in the following extracts:

(9) <B2> em well em we visit three countries and erm . it's a shame that you: didn't come . it was beautiful really <\B2>
(Role-Play Learner Corpus)

(10) <B2> yes em we call also the other . boy ... Peter <\B2>
(Role-Play Learner Corpus)

What is going on here is that the speakers do not seem to have realized that *BEAUTIFUL* and *BOY* are not semantically coextensive with Italian *BELLO* and *RAGAZZO*. Indeed, unlike Italian *BELLO*, *BEAUTIFUL* is not usually used to express a generic positive appreciation of an event (such as a trip – *NICE* would have been a better choice). By the same token, *BOY* does not generally refer to young adults, while in Italian a 35-year-old man may well be referred to as *RAGAZZO*. Other instances of lesser-known errors that are brought up by the Role-Play Learner Corpus originate from speakers departing from pragmatic norms – we have already provided extracts of role-plays where speakers appear to contravene Grice's Cooperative Principle.

From what we have been saying in this section it is evident that even a small learner corpus such as the one we have been compiling may reorient both researchers' and teachers' perspectives on learner language and learning/teaching priorities. The conclusion of the paper will look in more detail at the implications that learner corpora may have for language learning/teaching and teacher development.

6. Conclusion

The analysis of the Role-Play Learner Corpus carried out so far has enabled us to sketch out a tentative description of our students' spoken language and has singled out many of those errors that teachers of English generally and 'rather vaguely' regard as unacceptable, whilst at the same time highlighting errors of which EFL teachers are usually less aware. How can this kind of data be useful to both language teaching methodology and teacher development?

In language teaching, data from our spoken learner corpus can be used to encourage data-driven learning (Granger and Tribble 1998; Nesselhauf 2004). Teachers may choose some instances from the corpus in order to draw students' attention to some aspects of underuse, misuse or overuse of particular linguistic features, such as some pragmatic aspects of discourse management or lexicogrammatical features like those described above. This type of activity seems to be of particular relevance to advanced learners.

Focused negative evidence in data-driven learning is a good way to foster language acquisition but requires some parallel and follow-up activities – it is also necessary to provide positive evidence from comparable native speaker corpora. The teacher might start off by presenting some examples taken from the learner corpus and then show the native speaker usage or alternatively show the native speaker examples first. For example, using data from the Role-Play Learner Corpus, one could get students to replace *BEAUTIFUL* and *BOY* with more suitable words.

Another aspect related to the pedagogic use of learner corpora is their relevance to teacher development. Unfortunately, not many teachers are aware of the possibilities offered by corpus work. However, the use of corpora, and more specifically of learner corpora, would give them the opportunity to find out what "they have always wanted to know" (Tsui 2004).

This raises two main issues which should be tackled both in pre- and in-service teacher training. First of all, discovery activities such as those described above, encourage learners to become more autonomous. This requires "a supportive, non authoritarian environment" in which the teacher acts as "a learning expert rather than a language expert". Thus, discovery learning both with native and learner corpus data is "not only empowering for learners, but for teachers as well", and not only for "non-native speaker teachers" (Bernardini 2004: 28) but, we would argue, for native speaker teachers too.

Secondly, teachers need specific training in the use of concordancers in the classroom, in order to become adept at using the corpus as a reference tool and to find out about specific strategies involved in the heuristic process of analysing authentic data (Meunier 2002: 136). So far teachers have been able to rely on tools available online[16] but more recently ambitious research projects such as TeleNex (Allan 2002; Tsui 2004) have been set up to provide learners and teachers with integrated resources such as concordancing tools and lexical databases in CALL environment. Teachers' questions are answered using corpus data, including a learner corpus of primary students' written English and a corpus of secondary students' written and spoken English. The TeleNex project demonstrates how

> empirical linguistic data which show the context and frequency of occurrence of [specific] linguistic items can be a powerful tool to raise teachers' linguistic sensitivity, to help teachers question long-standing assumptions and to gain new insights into language structure and use. (Tsui 2004: 42)

We hope that our research based on the Role-Play Learner Corpus will make a useful contribution in this direction.

References

Allan, Q.G. (2002) "The TELEC secondary learner corpus: a resource for teacher development", in S. Granger , J. Hung and S. Petch-Tyson (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, John Benjamins, Amsterdam/Philadelphia, pp. 195-212.

Belz, J. (2004) "Learner corpus analysis and the development of language proficiency", *System* 32, pp. 577-591.

Belz, J. and N. Vyatkina (2005) "Learner corpus analysis and the development of L2 pragmatic competence in networked intercultural language study: the case of German modal particles", *Canadian Modern Language Review/Revue canadienne des langues vivantes* 62 (1), pp. 7-48.

Bernardini, S. (2004) "Corpora in the classroom: an overview and some reflections on future developments", in J. Sinclair (ed.), *How to Use Corpora in Language Teaching,* John Benjamins, Amsterdam/Philadelphia, pp. 15-36.

[16] Other resources for teachers can be found on Tom Cobb's website http://www.lextutor.ca/multi_conc/, which offers a tool to prepare exercises on different language items using concordances. An alternative resource is Tim Johns' Virtual Data-Driven Learning Library containing samples of concordance-based teaching and learning materials.

Black, M. (2007) “Objective IELTS: meeting the challenges of IELTS”, in B. Beaven (ed.), *IATEFL 2006. Harrogate Conference Selections*, IATEFL, Canterbury, pp. 180-182.

Bygate, M., P. Skehan and M. Swain (eds) (2001) *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*, Longman, London.

Carter, R. A. and M.J. McCarthy (1997) *Exploring Spoken English*, Cambridge University Press, Cambridge.

Carter, R.A. and M.J. McCarthy (2003) “If you ever hear a native speaker, please let us know!”, in A. Pulverness (ed.), *IATEFL 2003 Brighton Conference Selections,* IATEFL, Canterbury, pp. 116-123.

Corder, S.P. (1976) *Error Analysis and Interlanguage*, Oxford University Press, Oxford.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge University Press, Cambridge.

Dagneaux, E., S. Denness and S. Granger (1998) “Computer-aided error analysis”, *System* 26 (2), pp. 163-174.

Ellis, R. (1997) *Second Language Acquisition*, Oxford University Press, Oxford.

Ellis, R. (2003) *Task-based Language Learning and Teaching*, Oxford University Press, Oxford.

Ellis, R. and G. Barkhuizen (2005) *Analysing Learner Language*, Oxford University Press, Oxford.

Granger, S. (1994) “The learner corpus: a revolution in applied linguistics” *English Today* 39 (10/3), pp. 25-29.

Granger, S. (1998) “The computerized learner corpus: a versatile new source of data for SLA research”, in S. Granger (ed.), *Learner English on Computer*, Longman, London/New York, pp. 3-18.

Granger, S. (2002) “A bird's-eye view of computer learner corpus research”, in S. Granger, J. Hung and S. Petch-Tyson (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, John Benjamins, Amsterdam/Philadelphia, pp. 3-33.

Granger, S. (2003) “The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research”, *TESOL Quarterly* 37 (3), pp. 538-546.

Granger S., E. Dagneaux and F. Meunier (2002) *The International Corpus of Learner English, ICLE, Handbook*, with CD-ROM, UCL, Presses Universitaires de Louvain, Louvain-la-Neuve.

Granger, S. and C. Tribble (1998) “Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning”, in S. Granger (ed.), *Learner English on Computer*, Longman, London/New York, pp. 199-209.

Holliday, A. (1994) “Student culture and English language education: an international perspective”, *Language, Culture and Curriculum* 7 (2), pp. 125-143.

Mauranen, A. (2003) “The corpus of English as a lingua franca in international settings”, *TESOL Quarterly* 37 (2), pp. 513-27.

Mauranen, A. (2004) “Spoken corpus for an ordinary learner”, in J. Sinclair (ed.), *How to Use Corpora in Language Teaching*, John Benjamins, Amsterdam/Philadelphia, pp. 89-105.

McCarthy, M. (1998) *Spoken Language and Applied Linguistics*, Cambridge University Press, Cambridge.

Meunier, F. (2002) “The role of learner and native corpora in grammar teaching”, in S. Granger, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, John Benjamins, Amsterdam/Philadelphia, pp. 119-142.

Myles, F. (2005a) “Interlanguage corpora and SLA research”, *Second Language Research* 21 (4), pp. 373-391.

Myles, F. (2005b) “The emergence of morpho-syntactic structure in French L2”, in J.-M. Dewaele (ed.), *Focus on French as a foreign language: multidisciplinary approaches,* Multilingual Matters, Clevedon, pp. 120-135.

Nava, A. (2005) “Comparing tails. An exploratory study of tails in native spoken English and in Italian EFL learners’ interlanguage*”, Mots Palabras Words* 6, pp. 71-92.

Nesselhauf, N. (2004) “Learner corpora and their potential for language teaching”, in J. Sinclair (ed.), *How to Use Corpora in Language Teaching*, John Benjamins, Amsterdam/Philadelphia, pp. 125-152.

Nunan, D. (2004) *Task-based Language Teaching*, Cambridge University Press, Cambridge.

O’Keeffe, A. and M. McCarthy (2007) *From Corpus to Classroom: Language Use and Language Teaching,* Cambridge University Press, Cambridge.

Oller, J.W. Jr. (1979) *Language Tests at School*, Longman, London.

Prodromou L. (2003) “In search of SUE: the successful user of English”, *Modern English Teacher* 12 (2), pp. 5-13.

Pulcini, V. (2004) “A corpus of ‘informal academic interviews’: the Italian component of the LINDSEI project”, in M.T. Prat Zagrebelsky (ed.), *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners*, Edizioni dell’Orso, Alessandria.

Pulcini, V. and A. Damascelli (2005) “A corpus-based study of the discourse marker okay”, in A. Bertacca (ed.) *Historical Linguistic Studies of Spoken English*, Edizioni Plus, Pisa, pp. 231-243.

Pulcini, V. and C. Furiassi (2004) “Spoken interaction and discourse markers in a corpus of learner English”, in A. Partington, J. Morley and L. Harmann (eds), *Corpora and Discourse*, Peter Lang, Bern.

Reder, S., K. Harris and K. Setzler (2003) “The multimedia adult learner corpus”, *TESOL Quarterly* 37 (3), pp. 546-557.

Rizzardi, M.C., C. Degano, A. Nava and L. Pedrazzini (2004) "The language of role-play", paper presented at the 25th *ICAME* Conference, University of Verona, May 2004.

Rule, S. and E. Marsden (2006) "The acquisition of functional categories in early French second language grammars: the use of finite and non-finite verbs in negative contexts", *Second Language Research* 4 (22), pp. 188 - 218.

Seidlhofer, B. (2001) "Making the case for a corpus of English as a lingua franca", in G. Aston and L. Burnard (eds), *Corpora in the Description and Teaching of English*, Cooperativa Libraria Univeristaria Editrice, Bologna, pp. 70-85.

Seidlhofer, B. (2005) "English as a lingua franca", *ELT Journal* 59, pp. 339-341.

Sinclair, J. (1996) *EAGLES. Preliminary Recommendations on Corpus Typology*. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Skehan, P. (1998) *A Cognitive Approach to Language Learning*, Oxford University Press, Oxford.

Tan, M. (2005) "Authentic language or language errors? Lessons from a learner corpus", *ELT Journal* 59 (2), pp. 126-134.

Tsui, A. (2004) "What teachers have always wanted to know – and how corpora can help", in J. Sinclair (ed.), *How to Use Corpora in Language Teaching,* John Benjamins, Amsterdam/Philadelphia, pp. 39-61.

Willis, D. and J. Willis (2007) *Doing Task-based Teaching,* Oxford University Press, Oxford.

6. Corpora and Historical Studies

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 401-412 © 2008 Polimetrica International Scientific Publisher Monza/Italy

19CSC, second-generation corpora and the history of English

Marina Dossena and Richard Dury
University of Bergamo

1. Introduction

In the last few decades, Corpus Linguistics has had a tremendous impact on research in Italian universities, not least thanks to the work of colleagues like Maria Teresa Prat Zagrebelsky, to whom this collection is gratefully dedicated, who have established important and highly meaningful connections with work being done in other European institutions. Our contribution intends to highlight its indebtedness to this methodology and the implications it clearly has on both theoretical approaches and practical findings.

Corpus linguistics seems to have fairly deep roots now, almost half a century since the publication of such a ground-breaking collection as the Brown Corpus,[1] followed a few years later by the LOB Corpus,[2] and – as far as historical varieties were concerned – the Helsinki Corpus.[3] However, such foundations have never been considered to be an end in themselves, but the starting point for the advancement of investigations in a variety of aspects. As new corpora were compiled and published, new search tools were developed, and new potentialities uncovered, scholars realized how flexible and – at the same time – securely reliable corpus studies could be.

[1] See http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM.
[2] See http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM.
[3] See http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM.

In recent years, several research groups, especially at the University of Helsinki, have developed what is described as 'second-generation' corpora, i.e., collections of texts that do not necessarily encompass different registers and may serve multiple purposes, such as the early corpora mentioned before, but are compiled with specific research questions in mind, and may thus focus on a certain register or variety. As far as diachronic interests are concerned, this is the case, for instance, of the Corpus of Early English Correspondence Sampler (CEECS),[4] the Zurich English Newspaper Corpus (ZEN)[5] and the Corpus of Middle English Medical Texts (MEMT: see Taavitsainen *et al*. 2005).

In addition, increasing attention has been given to Late Modern English on the one hand, and to Scotland on the other. While the corpora currently being compiled in Edinburgh and Helsinki will comprise texts up to the end of the eighteenth century (Meurman-Solin 1999 and 2001; Williamson 2000), other projects on the nineteenth century restrict themselves to English in England (see Kytö *et al*. 2006).

Our project thus places itself at the intersection of complementary research interests, in that it intends to focus on usage in nineteenth-century Scotland. In particular, we intend to take into consideration authentic usage in correspondence, both formal and informal. The reasons behind this decision are numerous. First of all, the study of correspondence allows us to investigate a range of different text types within the same category, from business and official correspondence to familiar letters. This means we can take into consideration documents encoded by users whose linguistic competence may vary greatly depending on their level of schooling and circumstances; we may come across letters that follow highly codified models of business writing and texts that are actual 'conversations in writing', in which the echo of spoken language is reflected in spelling, syntax and vocabulary. As a result, we have a series of texts allowing us to study a range of linguistic elements on the basis of meaningful sources. In the next sections the rationale of our work and some preliminary findings will be outlined.

[4] See http://khnt.hit.uib.no/icame/manuals/ceecs/INDEX.HTM.

[5] See http://es-zen.unizh.ch/.

2. 19CSC: method and data

At the time of writing (summer 2007), 19CSC comprises ca. 400 letters, for a total of ca. 100,000 orthographic units. Clearly, this is far from representative of nineteenth-century epistolary discourse, and much more work will be needed to achieve that aim. For now, however, we think it is important to stress a few key points concerning our method and some terminological issues.

First of all, the texts included in the corpus derive from the transcription of authentic, previously unedited, manuscripts[6] selected both randomly and in connection with each other (e.g., in the case of answers to enquiries, or rejoinders to controversies). As for encoders and recipients, they are men and women of varying ages, living in Scotland or of Scottish origin, whose level of education ranges from fully to minimally-schooled (see Fairman 2000, 2003). This allows us to access a variety of writing styles, more or less approximating the standards set in schools or indeed in letter-writing manuals aimed at self-help, and investigate the extent to which authentic usage could diverge from what would be expected, if nothing but printed materials were taken into consideration.

As for terminology, it may be useful to clarify that by ‘encoder’ we mean the person that appears to have written the letter; in the case of autographs, it is the same as the writer; when the letter was written by an amanuensis, instead, the encoder is the person who dictated or prompted the text (an illiterate speaker or, in fact, at the opposite end of the social scale, a highly educated manager who could afford a secretary). Similarly, we use ‘recipient’, instead of ‘addressee’, because it was not unusual for letters to be circulated and made accessible to a wider network of people than just the person whose address was provided on the verso of the sheet or, later, on the envelope. Consequently, we cannot assume the message to have been exclusively for anyone specific, unless this is indicated in the text itself. In familiar letters, for instance, the reference to greetings sent by or being sent to third parties is a clear

[6] We gratefully acknowledge permission to access and quote from MSS held in the National Library of Scotland, Glasgow University Archives, and the Bank of Scotland Archives in Edinburgh. Such permission does not extend to third parties, so the quotations presented in this paper should not be used elsewhere.

indication that the letter was not meant to be kept secret, and its contents could perhaps be summarized, if not actually circulated, to such external participants in the exchange. In practice, the letter cannot be assumed to have been strictly private in the current sense of the word. Similarly, business letters could include references to ‘clients’, ‘legal representatives’, or ‘directors’, who had a say in the topic under discussion, and who could therefore be expected to have access to the contents of the letters, as this was a matter of their concern.

It will have been noted that we described the size of the corpus in terms of ‘orthographic units’, and not words or tokens, because in nineteenth-century handwritten documents words are often joined up by a continuous line of the pen and ‘additionally’ juxtaposed. This second case is often found with “I am”, but other examples are given below:

(1) it reminds me of the good old times when Iused to here the Gaelic song in the Bonnie Highlands

(2) Iintend this year tostick in + make a nest egg to send home tho’ as to farming there on a small scale Ihave my doubts of.

(3) I think the parties are much indebted toyou forthe friendly manner in which youhave acted

While it might be supposed that this phenomenon occurs in letters by minimally-educated users, because in those it is more frequent to record instances of phonetic spelling, it is actually significant to see that linked-up words also occur quite frequently in documents encoded in elegant copperplate, secretarial hands – this is the case, for instance, of (3) above. The people who used these hands, clearly trained in formal lettering and fairly educated, at least in a professional domain, could therefore use this device to indicate syntactic closeness of elements (as in the frequent attachment of the subject pronoun to the verb). Decorated letters (for example a final-d looped forward then back and under the word) could have both syntactic and pragmatic functions (see Dury 2006 and forthcoming). From the point of view of the corpus linguist, however, the presence of linked-up words in a manuscript creates a problem if a study of the type/token ratio is envisaged. In this case, the transcription with the exact representation of the text should be preserved, and another copy of the transcription should be made, in which tokens are separated, and provided with a tag that allows the researcher to identify the artificial division, – e.g.

(4) I think the parties are much indebted to <_> you for <_> the friendly manner in which you <_> have acted[7]

Similarly, tags are added to indicate superscript forms and self-corrections. In the quotation below, for instance, we have the indication of superscript (^^) and of self-correction by means of crossed out letters:

(5) will come Back when the water goes down ^in July^ and go to work again untill fall ~~^in July^~~ perhaps at 6 $ Pr day

It is true that self-corrections may be difficult to read, but when they are both visible and intelligible, they should be recorded as accurately as possible (when the correction is inked out completely, a transcriber's note should be added to that effect, as when a word is illegible because the page is torn, or the mark of the seal has hidden it). Of course, self-corrections occur more frequently in documents written by encoders who could not afford to write a draft (let alone more than one) and a fair copy. However, they may occur to repair oversights in copies of formal letters, and of course are observed in drafts of letters written by high status encoders (though such texts were not normally preserved, and are therefore more difficult to collect and study).[8]

[7] Note that the tag has to be between angular brackets, lest in any automatic search it is confused with a punctuation mark, i.e. the line that could follow the full stop in nineteenth-century MSS, but which fell out of use in the twentieth century.

[8] Civil service exams for clerical posts at the end of the nineteenth century and early twentieth century included an exercise in making a fair copy of an untidy facsimile draft with crossings-out, insertions and abbreviations (see Anon. 1889; Jones 1912). This means that "I am &c." used for the draft, had to be expanded in the copy, using the appropriate formula.

Overview

Such considerations shed light on the complexity of studies in letter-writing. Far from being (simple?) ‘conversation in writing’, these texts offer insights into register variation, and their analysis may profit from the application of the methodological tenets of social network theory. This means that quantitative studies may be significant, but need to be interpreted on the basis of accurate overviews based on the principles of social network theory. In addition, the history of specialized discourse – of business dynamics, in particular – may be employed to investigate variation and change in documents that were expected to reflect the pragmatic models of the time, with ideas of style, politeness, and indirectness that do not necessarily mirror those observed in present-day discourse. Indeed, the cultural traits of nineteenth-century discourse communities need to be taken into consideration for the investigation of the strategies employed to convey stance, authority, evaluation, or judgement. As changes in scientific thought styles have been shown to have had an impact on scientific discourse in Early Modern times (Taavitsainen 2001), we can assume that the economic, industrial and cultural background may have influenced the characteristic of written business discourse in Late Modern times. This means that the contribution of social and business historians cannot be neglected, if we are to draw reliable conclusions from the study of these documents.

As for familiar letters, these include a considerable number of letters (about 1/6 of the whole corpus) written to friends and relatives by people who had emigrated from Scotland to the USA, Canada, Australia and New Zealand. Through them we can also study the way in which point of view is conveyed, as their descriptions are seldom neutral, but typically convey their opinions about what they see (whether it is a natural or human environment), the circumstances in which they find themselves, and their expectations about life abroad and in the home country. In addition, it is particularly promising to consider the possibility of studying these in connection with letters written by other emigrants in different languages (e.g. cf. Elspaß 2002, on the letters of German emigrants to the USA), in order to study formulaic usage and pragmatic choices in a cross-linguistic perspective.

A. Martelli and V. Pulcini (eds) *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*

3. Preliminary findings

As the corpus cannot be assumed to be representative yet, it would be inappropriate to offer quantitative findings based on it, as their statistical significance would be extremely limited. However, some preliminary, albeit still very tentative, observations may be offered from the qualitative point of view.

First of all, it is very interesting to see that different discourse communities and social networks interact quite dynamically in these documents. Such networks and communities are observed to change and adopt different communicative strategies as social or geographical distance increases, or mutual acquaintance becomes more firmly established. For instance, we have instances of women writing to, or receiving letters from, their families of origin after marriage. Fairly often, the sense of belonging to a new family is signalled through news pertaining to the new (hopefully happy) reality of the bride. This, however, does not exclude references to how much one correspondent misses the others, and to recollections of time spent together. An example is given below:

(6) Dear Kitty,
We miss you very much
It was so lonely in bed without
you, that I got your pillow into
bed to hug, but it was not nearly
so nice as yourself. I did not go
to school on Thursday, because I could
not prepare any lessons on Wednesday
evening. […] We ate such a lot of Mar-
riage cake on Wednesday, but
we were not at all the worse
of it. I hope you are enjoying
yourself. […] Grant + Agnes
+ Walter send~~end~~ their
very dear love to you; and also
to Mr Oliver. […] Please
exuse the bad writing of this.
Walter joins with me in hop-
ing that neither of you will be
sick in crossing. Please write

to me soon, if you have time.
Give my love to Mr Oliver, and
with a very great deal to yourself,
~~I am~~,
your very
I am,
your very loving sister, […]

In a similar framework, letters encoded by emigrants display intense psychological proximity in spite of geographical distance. Recipients are indirectly invited to share the experiences of the encoders through vivid descriptions that stress the encoder's point of view and invite solidarity, sympathy, or appreciation. In this way, important social bonds are maintained in spite of separation (see Dossena 2007 and forthcoming). While the formulaic character of letters that attempted to imitate fully-schooled models transpires and is indeed very obvious at the beginning and at the end of the letter, throughout the text we observe that overt dialect forms are avoided. Less marked, or perhaps less frequently stigmatized forms, however, do occur, allowing scholars access to instances of vernacular syntax and morphology. The example below illustrates this mixture of formulaic and personalized language:

(7) Dear Brother
I received yours letter last night and
was glad to here of you all being in
good health as this leaves me at present
hoping this will find you the same.
I see that you have stormy weather since
a while at home but I was glad
to here that you got down your
crop. […] there is hardly any
oats to be seen
at all Indian corn
and flax and such like […]
the water is very good indeed the longer
I am in the place the more I like it
any person could not desire for a more
beautiful place to stop in, […]
I see you did not commence in
the fishing yet when you will

write me let me know how matters
stands over there in the way of
work and people write soon I must
close at present   Yours sincerely Dear
                        Brother

In business exchanges, different types of letters, reflecting different types of business relationships, are recorded. On the one hand, we have fairly distant, official circulars issued to agents and representatives, in which the level of personalization is minimized. On the other, we have highly personalized letters that, while still dealing with business issues, address someone specifically and, as a result, the strategies selected by the encoder reflect the changes and developments in the relationship. Especially in cases of conflict, it is interesting to see how negotiation is conducted; in particular, in these cases, in which directions, requests, complaints, and other face-threatening speech acts are not uncommon, authority can be conveyed more or less forcefully, with more or less deference to the recipient. This means that positive and negative politeness strategies can be observed to be employed in different contexts, depending on what kind of status the encoder wishes the recipient to perceive. When positive politeness is employed, for instance, the encoder's status may be artificially downplayed, in order to raise the one of the recipient. As a result, we find instances of 'fluctuating identities', in which the pragmatic value of the proposition dictates the linguistic choices of the encoder and shapes the text according to the interpretation intended by the encoders themselves (see Dossena 2006a, 2006b).

4. Conclusion

At the turn of the millennium, English linguistics seems to be profiting considerably from the new methodology offered by corpus studies. New collections of texts, accompanied by increasingly sophisticated software for data retrieval and analysis, and the insights provided by discourse analytical and pragmatic frameworks of investigation in a diachronic perspective, allow new research trends to be outlined and pursued.

In the case of 19CSC, focus is on geo-historical variation in correspondence of different types. The range of manuscripts being transcribed and investigated allows us to benefit from the theoretical approaches outlined by scholars whose work has always relied extensively on authentic manuscripts, i.e. Old and Middle English dialectologists. They are well aware of the importance of recording and preserving as much information as possible about the text and its (even internal) variants, but also of the virtually unending range of information that could be extracted from the text. At the same time, the relatively recent interest in Late Modern English prescriptivism, with its attempts to codify and stabilize the English language, has provided new information on textbooks, self-help manuals, and the attitudes to dialectal and sociolectal variation in different parts of the country. As a result, it is now possible to study our manuscripts within a fairly accurate socio-historical framework that allows us to identify the most significant models that encoders attempted to imitate, and to assess the extent to which there were instances of divergence, dictated by the need for originality which is indispensable in personalized, non-stereotypical communication

Undoubtedly, research will progress both in relation to methodological tools and theories, and in relation to data and findings. The combination of these two aspects will lead to further knowledge about the linguistic strategies, tenets and cultural frameworks in which discourse communities of the past found themselves to operate. If it is true that “the past is a foreign country”, relatively small, but clearly focused, second-generation corpora can contribute to its mapping in a very significant way.

References

Anon. [G.L. Dunnet] (1889) *Military Handwriting and the Copying of Official Manuscripts, by an Army Schoolmaster*, Gale & Polden, Chatham.

Dossena, M. (2006a) “Stance and authority in nineteenth-century bank correspondence – a case study”, in M. Dossena and S.M. Fitzmaurice (eds), *Business and Official Correspondence: Historical Investigations*, Peter Lang, Bern, pp. 175-192.

Dossena, M. (2006b) “Forms of self-representation in 19th-century business letters”, in M. Dossena and I. Taavitsainen (eds), *Diachronic Perspectives on Domain-Specific English*, Peter Lang, Bern, pp. 173-190.

Dossena, M. (2007) “As this leaves me at present” – Formulaic usage, politeness and social proximity in nineteenth-century Scottish emigrants’ letters”, in S. Elspaß, N. Langer, J. Scharloth and W. Vandenbussche (eds), *Germanic Language Histories from Below (1700-2000)*, Walter De Gruyter, Berlin, pp. 13-29.

Dossena, M. (forthcoming) “‘Many strange and peculiar affairs’: point of view, stance and evaluation in Scottish emigrants’ letters of the 19th Century”, *Scottish Language*.

Dury, R. (2006) “A corpus of nineteenth-century business correspondence: methodology of transcription”, in M. Dossena and S.M. Fitzmaurice (eds), *Business and Official Correspondence: Historical Investigations*, Peter Lang, Bern, pp. 193-205.

Dury, R., forthcoming, “The history of the copperplate hand and the importance of noting handwriting styles in letter transcriptions”, in M. Dossena and I. Tieken-Boon van Ostade (eds), *Studies in Late Modern English Correspondence: Methodology and Data*, Peter Lang, Bern.

Elspaß, S. (2002) “Standard German in the nineteenth century? (Counter)Evidence from the private correspondence of ‘ordinary people’”, in A.R. Linn and N. McLelland (eds), *Standardization – Studies from the Germanic Languages*, John Benjamins, Amsterdam, pp. 43-65.

Fairman, T. (2000) “English pauper letters 1800-34 and the English language”, in D. Barton and N. Hall (eds), *Letter Writing as a Social Practice*, John Benjamins, Amsterdam, pp. 63-82.

Fairman, T. (2003) “Letters of the English labouring classes 1800-34 and the English language”, in M. Dossena and C. Jones (eds), *Insights into Late Modern English*, Peter Lang, Bern, pp. 265-282.

Jones, A.J.L. (1912) *Pitman’s Guide to Candidates for His Majesty’s Civil Service in Copying Manuscript, Autography, Handwriting, etc. Actual Examination Papers Only*, Pitman, London.

Kytö, M., M. Rydén and E. Smitterberg (eds) (2006) *Nineteenth-century English: Stability and Change*, Cambridge University Press, Cambridge.

Meurman-Solin, A. (1999) “Letters as a source of data for reconstructing Early Spoken Scots”, in I. Taavitsainen, G. Melchers and P. Pahta (eds), *Writing in Non-standard English*. John Benjamins, Amsterdam, pp. 305-322.

Meurman-Solin, A. (2001) “Women as informants in the reconstruction of geographically and socio-culturally conditioned language variation and change in 16th and 17th century Scots”, *Scottish Language* 20, pp. 20-46.

Taavitsainen, I. (2001) “Evidentiality and scientific thought-styles: English medical writing in late Middle English and Early Modern English”, in M. Gotti and M. Dossena (eds), *Modality in Specialized Texts*, Peter Lang, Bern, pp. 21-52.

Taavitsainen, I., P. Pahta and M. Mäkinen (compilers) (2005) *Middle English Medical Texts*, John Benjamins, Amsterdam.

Williamson, K. (2000) “Lexico-grammatical tags and the phonetic and syntactic analysis of medieval texts”, in C. Mair and M. Hundt (eds), *Corpus Linguistics and Linguistic Theory, Language and Computers*, Rodopi, Amsterdam, pp. 385-395.

A. Martelli and V. Pulcini (eds), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, 413-422 © 2008 Polimetrica International Scientific Publisher Monza/Italy

Language change and variation in English: the case of unsplit *FOR TO* in infinitival purpose clauses

Gerardo Mazzaferro – University of Turin

1. Introduction

The aim of this paper is to offer a corpus-based investigation of unsplit *FOR TO* in infinitival purpose clauses. Drawing on both diachronic and synchronic corpora of English we will attempt to: describe first the general properties of infinitival clauses in Present-day English (henceforth PDE), and, second, consider the origins and development of *FOR TO* from the Middle to the early Modern English periods (1150-1700) (henceforth ME and eModEngl).

We will show that although *FOR TO* is not a feature of standard PDE it is still part of the grammar of almost all English regional dialects and it may also be found in some non-native varieties of English, learner interlanguages and second language acquisition contexts.[1] According to Pak (2005: 1):

> "[I]n standard modern English, an infinitive can only be marked with *for* if it also contains a overt subject, so that sentences like (1a) are ruled out. Furthermore, *for* is obligatory in contexts like (1b), where the infinitive is itself a sentential subject (c.f. 1c).
>
> *(1)Standard modern English*:
>
> a. *John went to the store for to buy bred.
> b. For Mary to travel so far is no small thing.
> c. *Mary to travel so far is no small thing […]".

[1] The examples below are taken from both the International Corpus of English (Greenbaum 1996) and the International Corpus of Learner English (Granger, Dagneaux and Meunier 2002).

2. The data

A variety of corpora have been explored in order to find evidence of the use of unsplit *FOR TO* in infinitival purpose clauses, particularly: 1) The Helsinki Corpus of English Texts (Kytö 1991), which contains a diachronic part covering the three main periods of the English language, namely Old English (700-1150), Middle English (1150-1500) and early Modern English (1500-1710c.) and it is of about 1.700,000 words; (2) and The Freiburg Corpus of English Dialects Sampler, which consists of 121 transcribed interviews (1970-2000) covering five main dialectal areas: the southwest, southeast, Midlands, north and Scottish Lowlands of England and it contains almost 1.000,000 words (Szmrecsanyi and Hernández 2007).

Lastly, on the methodological level our study represents a quantitative analysis of *FOR TO*, which aims at finding frequencies of this particular grammatical feature at different stages of its development and use.

3. Present-day English infinitival clauses

In PDE we distinguish between finite and non-finite clauses: the former always contain a subject and predicate; while the latter can be constructed without a subject:

> "In languages generally, the concept of finiteness normally applies in the first instance to verbs, and then derivatively to clauses, a clause being finite or non-finite according as the verb is finite or non-finite […]. A finite verb is one that is inflected for person and number […]. [Further] finite clauses may be main or subordinate but non-finite ones are always subordinate […]" (Huddleston and Pullum 2006: 208).

There are four main classes of non-finite clauses: 'to-infinitival', 'infinitival without to' or 'bare infinitival', '-ing or gerund participial', and '-ed or past participial' (Quirk and Greenbaum 1973: 310-311); e.g.

(1) I want to repaint the kitchen [to infinitival]

(2) I helped repaint the kitchen [bare infinitival]

(3) Inviting the smiths was a mistake [gerund participial]

(4) Their son was among those arrested for drunkenness [past participial] (Huddleston and Pullum 2006: 215)

In PDE 'to-infinitival' clauses, which contain an overt subject, are commonly introduced by *FOR*; e.g.

(5) for him to go to the seaside is very relaxing.

"[W]e take this *FOR* to be a clause subordinator, comparable to the that of finite clauses […]" (Huddleston and Pullum 2002: 65).

4. Present-day English infinitival purpose clauses

PDE purpose clauses are generally infinitival adjuncts:

> "[…] [s]yntactically, they can be preceded by *in order* and characteristically can be moved to front position: […] i *He walked [(in order) to save money]* [;] ii *[(in order) to save money, he walked]* [italics mine] (Huddleston and Pullum 2002: 1222).

SO AS is also used to express purpose and goal as in:

(6) I left at 5 a.m. so as to get to Milan early.

Finite clauses of purpose are, on the contrary, expressed by *IN ORDER THAT* and *SO AS THAT*, like for example:

(7) John visited London in order that [so as that] he could see his MP (Quirk and Greenbaum 1973: 328).

5. An historical overview

Old English

The use of infinitives to express a complement is common throughout the history of the English language. In Old English there are two kinds of infinitives: a 'bare or zero infinitive' [verb stem+endings *–AN* and *–IAN*] and an 'inflected' one or 'to-infinitive' [*TO*+verb stem+ending *–ENNE*] (Fisher, van Kemenade and van der Wurff 2000: 62). The former is commonly found after auxiliaries such as *CUNNAN*, *WILLAN*, *SCULAN MAGAN* and *MŌTAN*; while the latter is used to signal purpose, necessity and obligation like for example in:

(8) an wulf wearð asend to bewwrigenne þæt heafod […]
A wolf was sent to guard the head […]

(9) Is eac to witanne […]
It must also be noted […] (Mitchell and Robinson 1992: 112).

To-infinitive constructions to express purpose continued at least until the end of our period, when some changes occurred. These are probably due to the fact that:

> "*To*, in origin a directional adverb/preposition, started off as an indicator (mainly) of purpose, but by late Old English /early Middle English it had lost that function so that it began to occur where previously only the bare infinitive was found. It is very likely that *to* increased its territory because it became a useful sign of the infinitive form, to distinguish it from other forms of the verb. Due to the reduction and loss of inflections, the infinitival endings […] could no longer serve that purpose" (Fisher 1992: 317).

Middle English

At the beginning of the ME period *TO* probably lost its original meaning as it is attested by the introduction of *FOR TO* marked infinitives, see Table 1. below, to strengthen and express direction and purpose like in:

(10) Ne we ne beoð iboren for to habbene nane prudu ne forðe nane oðre rencas ah we beoð on þisse liue for to ernien þa eche blisse in houeneriche[…] (We are not born to have pride nor even any other vanities; but we are in this life that we may earn the bliss in the kingdom of heaven […] (Morris 1969, 1968: 7).

Corpus sub-periods	**For to p. infs.**	**Words**	**% 1,000**
ME I (1150-1250)	11	113,010	0.09
ME II (1250-1350)	87	97,480	0.9
ME III (1350-1420)	420	184,230	2.28
ME IV (1420-1500)	212	213,850	0.10
Total	730	608,570	3.37

Table 1. Overall distribution of *FOR TO* infinitival purpose clauses in ME.

However, it is only towards the end of our period that *FOR TO* started to be used: "[…] as an equivalent of the to infinitive, sometimes for reasons of metre and rhyme [as in]:

(11) ne wonde/þis aventure for to frayn […] (Burrow and Turville-Petre 1992: 48).

Early Modern English

FOR TO infinitival constructions are also attested throughout the eModEngl period (see Table 2 below) when it is used interchangeably with 'to infinitive'. The choice between the two is only, as has been recently stated by Rissanen (1999: 288), "a question of preference or style"; in other words *FOR TO* does not represent a marker of purpose and goal any longer and it starts to be substituted by PDE constructions *IN ORDER TO* and *SO AS TO*; e.g.

(12) […] shipped or put on any Boat or Vessell in order to be shipped for Exportation […]

(13) […] He was sorry he had lived so as to wast his strength so soon […]

Corpus sub-periods	For to p. clauses	Words	% 1,000
EModEngl I (1500-1570)	46	190,160	1.27
EModEngl II (1570-1640)	8	189,800	0.0042
EModEngl III (1640-1710)	4	171,040	0.023
Total	56	551,000	1.33

Table 2. Overall distribution of *FOR TO* infinitival purpose clauses in EModEngl.

Present-day use of FOR TO

As stated above, after the seventeenth century *FOR TO* continues though its use is linguistically, socio-culturally and geographically restricted (Danchev and Kytö 2003: 37). Its presence is nowadays mainly attested in the English regional dialects, particularly in the southwest, southeast and north (see Table 3 below):

> "For complement clauses, the most pervasive feature[s] in the British Isles [is] […] unsplit for to infinitival purpose clauses […] as in *there was always one man selected for to make the tea* […]. The only

> variety where unsplit for to is not attested is East Anglia. In ScE [Scottish English] the infinitive is regularly marked by for to also in non-purpose contexts like *You werenae allowed at this time for to go and take another job on* […] (Kortmann 2004: 1095).

Present-day British Dialects	For to p. clauses	Words	% 1,000
Southwest	43	264,863	0.16
Southeast	44	260,643	0.17
Midlands	1	152,535	0.006
North	24	266,955	0.09
Scottish Lowlands	1	66,400	0.015
Total	113	1.011,396	0.44

Table 3. Overall distribution of *FOR TO* in PDE English dialects.

ESL, EFL varieties and learner interlanguages

A limited number of occurrences may be found in both non-native, mainly ESL and EFL, varieties of English such as Indian or Philippines English:

(14) What is uh solution for to overcome this corruption […]

(15) There was no time for to speak with […]

(16) Women need it for to get along uhm through […]

(17) No I think you were identified for to teach curriculum design […];

and learner interlanguages, particularly Italian, Spanish, and French; e.g.:

(18) […] So for to be part of the Union has not been […] (Italian learner of English)
(19) […] it seems to be more than just a need for to do something […] (Spanish learner of English).

The latter may represent instances of L1 transfer. For example, in Italian prepositions *PER* (and *A*) (PDE *FOR, TO*) are commonly used to introduce purpose clauses (*proposizioni finali*) which are commonly translated in English by *IN ORDER TO* or *TO*; e.g.:

(20) *Bisogna credersi davvero belle per partecipare a Miss talia* (You have to think you are really beautiful (in order) to partecipate in Miss Italy) (Maiden and Robustelli 2000: 377).

6. Conclusion

In conclusion, our analysis has attempted to highlight the fact that the use of *FOR TO* has continued from the early Middle English period, in which it originated, until PDE. Historically, *FOR TO* is used in both ME and eModEngl in a wide range of contexts and across different textual typologies and its use reaches a peak in the 14th and 15th centuries, see Tables 4 and 5 below.

Textual Typology	***N* Words**	***N* Occurrences**	**% 1,000**
Documents	4,540	49	1.07
Law	11,240	-	-
Proceedings*	-	-	-
Handbooks	30,080	22	0.07
Philosophy	10,170	12	0.11
Homilies	50,040	54	0.10
Sermons	25,010	75	0.29
Religious Treaties	122,720	173	0.14
Rules	7,200	10	0.13
History	49,310	83	0.16
Biography/Lives	20,820	5	0.24
Fiction	21,520	69	0.32
Romances	53,450	73	0.13
Drama/Mystery Plays	20,100	38	
Bible	37,720	10	0.18
Priv. Letters	3,540	7	1.9
Non-priv. Letters	4,060	11	0.27
Diaries*	-	-	-
Travelogue*	-	-	-

Table 4. Overall distribution of *FOR TO* across different textual typology in ME.

Textual Typology	***N* Words**	***N* occurrences**	**% 1,000**
Documents*	-	-	-
Law	36,750	1	0.02
Proceedings	-	2	-
Handbooks	33,660	3	0.090
Philosophy	25,590	3	0.11
Homilies*	-	-	-
Sermons	27,270	7	0.25
Religious Treaties*	-	-	-
Rules*	-	-	-
History	33,170	5	0.15
Biography/Lives	32,120	6	0.18
Fiction	6,080	-	-
Romances*	-	-	-
Drama/Comedies	22,380	3	0.13
Bible	43,420	7	0.16
Priv. Letters	3,670	4	1.08
Non-priv. Letters	2,150	1	0.46
Diaries	44,030	4	0.09
Travelogue	26,180	10	0.38

Table 5. Overall distribution of *FOR TO* across different textual typology eModEngl.

However, there are two main positions about the origins of *FOR TO*. The first refers to Pak (2005), who has recently investigated the distinction between *FOR* in subjectless *FOR TO* infinitives and the *FOR* in infinitives with subjects. She has convincely demonstrated that:

> " […] infinitival *for* is not uniformly a complementizer across English dialects, but some dialects have ineherited from western ME [Middle English] a single head *for to* that alternates with *to*".

The second points out, according to Danchev and Kytö (2003), that the development of *FOR TO* is probably due to external factors, namely the Scandinavian and French or Anglo-Norman presence in

England. *FOR TO* infinitive constructions are mostly preserved in areas which were subject to Scandinavian influence and they may have been borrowed or translated directly from it; see for example the Danish constructions *FOR AT* and *FOR TILL.*

Alternatively, ME *FOR TO* may derive from French or Anglo-Norman *FOR*+bare infinitive forms such as:

(21) þat he were mid heom ilome For teche heom of his wisdoms:

> "[...] *for* could have later spread to the *to* infinitive thus providing the needed new emphasis of the purposive meaning. [...] The probability of such a scenario is supported by quite a few other cases of French prepositional usage transfer to Middle English, e.g. in phrases such as by sea and by land (from *par mer* and *par terre*) [...]" (Danchev and Kytö 2003: 38-39).

References

Burrow, J.A. and T. Turville-Petre (1992) *A Book of Middle English*, Blackwell, Oxford/Cambridge (U.S.A.).

Danchev, A. and M. Kytö (2003) "The Middle English 'for to + infinitive' construction: a twofold contact phenomenon", in D. Kastovsky and A. Mettinger, *Language Contact in the History of English*, Peter Lang, Frankfurt, pp. 35-55.

Fisher, O. (1992) "Syntax", in N. Blake (ed.), *The Cambridge History of the English Language*, Vol. II , Cambridge University Press, Cambridge, pp. 207- 408.

Fisher, O., A. van Kemenade and W. van der Wurff (2000) *The Syntax of Early English*, Cambridge University Press, Cambridge.

Fisher, O. and W. van der Wurff (2006) "Syntax", in R. Hogg and D. Denison (eds), *A History of the English Language*, Cambridge University Press, Cambridge, pp. 109-198.

Granger S., E. Dagneaux and F. Meunier (eds) (2002) *International Corpus of Learner English*, UCL Presses Universitaires de Louvain, Louvain.

Greenbaum, S. (1996) *Comparing English Worldwide. The International Corpus of English*, Clarendon Press, Oxford.

Huddleston, R. and G.K. Pullum (2002) *The Cambridge Grammar of the English Language*, Cambridge University Press, Cambridge.

Huddleston, R. and G.K. Pullum (2006) "Coordination and subordination", in B. Aarts and A. McMahon (eds), *The Handbook of English Linguistics*, Blackwell, Malden/Oxford, pp. 220-243.

Kytö, M. (1991) *Manual to the Diachronic Part of Helsinki Corpus of English texts: Coding, Conventions and Lists of Source Texts*, Department of English, University of Helsinki, Helsinki.

Kortmann, B. (2004) "Synopsis: morphological and syntactic variation in the British Isles", in Kortmann B. and E.W. Schneider (eds.), *A Handbook of Varieties of English*", Mouton de Gruyter, Berlin/New York, pp. 1089-1103.

Martin, M. and C. Robustelli (2000) *A Reference Grammar of Modern Italian*, Arnold, London.

Mitchel, B. and F.C. Robinson (1992) *A Guide to Old English*, Blackwell, Oxford/Cambridge (U.S.A.).

Morris, R. [1868] (1969) *Old English Homilies and Homiletic Treaties (Sawles warde, and Pe wohunge of Ure Lauerd: Ureisuns of Ure Louerd and of Ure Lefdi, &c.) of the Twelfth and Thirteenth Centuries*, Greenwood Press Publishers, New York.

Pak, M. (2006) "Infinitive marking with for: a diachronic account", in *Penn Working Papers in Linguistics* 12 (1), pp. 293-306.

Quirk, R. and S. Greenbaum (1973) *A University Grammar of English*, Longman, Essex.

Rissanen, M. (1999) "Syntax", in R. Lass (ed.), *The Cambridge History of the English Language*, Vol. III, Cambridge University Press, Cambridge, pp. 187-331.

Szmrecsanyi, B. and N. Hernández (2007) *Manual of Information to Accompany the Freiburg Corpus of English Dialects Sampler (FRED).*
http:// www. freidok.uni-freiburg.de/volltexte/2859

Van Ek, J.A. and N.J. Robat (1984) *The Student's Grammar of English*, Basil Blackwell, Oxford.

Contributors

Simona Anselmi, Lecturer in English Language and Linguistics, Catholic University of Piacenza. Her research interests include the use of English as a lingua franca, contact linguistics and translation studies, with particular reference to self-translation, the use of translation in postcolonial contexts and corpus-based translation studies. Her recent publications include *La traduzione postcoloniale in Irlanda: 'Finnegans Wake', una traduzione in corso* (2005), "La traduzione nella scrittura postcoloniale" (2006) and "La traduttologia" (co-authored with Margherita Ulrych, 2008).

Guy Aston, Full Professor of English Language and Linguistics, SSLMIT (*Scuola Superiore Lingue Moderne per Interpreti e Traduttori*), University of Bologna, Forlì. His main fields of interest and research include conversational analysis, contrastive pragmatics, corpus linguistics and autonomous language learning. He was responsible for the PIXI project on the pragmatics of service encounters in English and in Italian, and collaborates with the University of Oxford on the British National Corpus project. His most important books are: *Learning Comity: An Approach to the Description and Pedagogy of Interactional Speech* (1988), *The BNC Handbook: Exploring the British National Corpus with SARA* (with Lou Burnard, 1998), *Learning with Corpora*. (edited, 2001), *Corpora and Language Learners* (edited with Silvia Bernardini and Dominic Stewart, 2004). Web page: http://www.sslmit.unibo.it/~guy/guypage.htm

Julia Bamford, Full Professor of English, *Facoltà di Lingue e Letterature Straniere*, University of Naples "L'Orientale". Her research interests lie in the field of spoken language with special reference to academic and business discourse from the point of view of genre analysis and evaluation. She is the author of *You Can Say That Again: Repetition in Discourse* (2000) and editor of *Evaluation in Oral and Written Academic Discourse* (with Laurie Anderson, 2004) and *Dialogue within Discourse Communities: Metadiscursive Perspectives on Academic Genres* (with Marina Bondi, 2005). Some of her most recent articles include "Symbolic and gestural uses of deixis in academic lectures" (2004) and "Subjective or objective evaluation?: Prediction in academic lectures" (2005).

Paul Bayley, Full Professor of English Linguistics, Facoltà di Scienze Politiche "Roberto Ruffilli", University of Bologna. His research combines corpus linguistics methodology, in particular through specialized corpora, with discourse analysis approaches informed by Systemic Functional Grammar. He is currently a participant in a VI Framework programme on European citizenship, IntUne (Integrated and United: A Quest for Citizenship in an ever closer Europe), involving political scientists and linguists working on a multilingual corpus of 150 million tokens. His recent publications include *Cross-cultural Perspectives on Parliamentary Discourse* (2004), "Terror in political discourse: from the Cold War to the unipolar world" (2007) and "Perhaps... but: expanding and contracting alternative viewpoints" (2007).
Web page: http://afferenti.sitlec.unibo.it/curriculum/ BayleyP.htm

Marina Bondi, Full Professor of English Language, *Facoltà di Lettere e Filosofia*, University of Modena and Reggio Emilia. She has published on various aspects of discourse analysis and EAP, with particular reference to the argumentative features of academic discourse and to the role of metadiscourse and evaluative language. Her recent work centres on language variation across genres, disciplines and cultures through the analysis of small specialized corpora. She has recently co-edited a volume on *Academic Discourse across Disciplines* (with Ken Hyland, 2006) and published papers on the specific features of historical writing ("Authority and expert voices in the discourse of history", 2007) and on the notion of "key-words" ("Key-words and emotions: a case study of the Bloody Sunday Inquiry", 2007).

Silvia Bruti, Associate Professor of English Language and Linguistics, *Facoltà di Lingue e Letterature Straniere*, University of Pisa. She has done research in text-linguistics, (historical) pragmatics and applied linguistics. Her publications are in the areas of cohesion and coherence, text complexity, applied linguistics ("The complexity of phoric relations: Anaphora *vs* cataphora", 2003). She has worked on reformulation and paraphrase, on which she edited a collection of essays (*La parafrasi tra messa a fuoco del codice e negoziazione discorsiva*, 2004). More recently her research has focused on intercultural pragmatics (English/Italian) and on the translation of pragmatic meaning in inter-linguistic subtitles and dubbing ("Cross-cultural pragmatics: the translation of implicit compliments in subtitles", 2006).
Web page: http://www.humnet.unipi.it/anglistica/curricula/bruti/

Sandra Campagna, Associate Professor of English Language and Linguistics, *Facoltà di Economia*, University of Turin. Her main research areas are translation studies, crosscultural studies (*Il tono comico: prospettive crossculturali* 1999) and sociolinguistics ("Voices from Bradistan: The Sense of Belonging in Multiracial Societies" 2001). Her recent research interests are in the fields of intercultural communication in economic discourse and in the web language domain (*Discoursal Strategies On-line: An Intercultural Approach to the Language of Charities*, 2004). She is currently working on identity issues and multimodal/multimedial discourse ("Website reading paths: constructing central and marginal identities in the semiotic landscape", 2007).

Umberto Capra, Researcher in English Language and Linguistics, *Facoltà di Lettere e Filosofia*, University of "Piemonte Orientale A. Avogadro", Vercelli. He chairs the Modern languages area of SIS-Piemonte and is chief editor of the *LEND* (*Lingua e Nuova Didattica*) journal. His main fields of research are language teaching, teacher education and the role of technologies in language teaching and learning. Among his publications: *Rainbow* (with Licia Bonzano, 1992), a video course of English for children; *Friendly* (1999) a bilingual English-Italian dictionary for schools, and *Tecnologie per l'apprendimento linguistico* (2005). Web page: http://www.lett.unipmn.it/%7Ecapra/capra_pubbl.htm

Stephen Coffey, Researcher in English Linguistics, *Facoltà di Scienze Politiche*, University of Pisa. His research interests focus on English lexicography, especially as regards monolingual learners' dictionaries, English phraseology, from a descriptive, pedagogical and comparative point of view (with Italian), high-frequency words of English, corpus linguistics and language learning. His recent publications include: "Considerations emerging from a frequency study of multiword units in a corpus of contemporary written Italian" (with Laura Cignoni, 2003), "High-frequency grammatical lexis in advanced-level English learners' dictionaries" (2006), "Investigating restricted semantic sets in a large general corpus: learning activities for students of English as a foreign language" (2007).

Giuseppina Cortese, Full Professor of English Language and Translation, *Facoltà di Scienze Politiche*, University of Turin. She has widely contributed to national and international journals, edited volumes and published essays/chapters on language education, domain-specific (academic) English, translation and gender. Her recent work is focussed on human rights discourse.

Her publications include*: Domain-specific English: Textual Practices across Communities and Classrooms* (co-edited with Philip Riley, 2002), *Identity, Community, Discourse. English in Intercultural Settings* (edited with Anna Duszak, 2005) and *Discourse Analysis and Contemporary Social Change* (co-edited with Norman Fairclough and Patrizia Ardizzone, 2007).

Marina Dossena, Full Professor of English Language, *Facoltà di Lingue e Letterature Straniere*, University of Bergamo. Her research interests focus on the features and origins of British varieties of English and the history of specialized discourse. In co-operation with Richard Dury, she is currently compiling a corpus of nineteenth-century Scottish correspondence. Recent publications include *Business and Official Correspondence: Historical Investigations* (co-edited with Susan M. Fitzmaurice, 2006), and *Diachronic Perspectives on Domain-specific English* (co-edited with Irma Taavitsainen, 2006). She is also the author of the monograph *Scotticisms in Grammar and Vocabulary* (2005).

Richard Dury, Associate Professor of English Language and Linguistics, *Facoltà di Scienze Umanistiche*, University of Bergamo. His research interests focus on the history of English in a European perspective and the works of Robert Louis Stevenson. Recent publications include: *Robert Louis Stevenson, Writer of Boundaries* (edited with Richard Ambrosini, 2006), "A corpus of nineteenth-century business correspondence: methodology of transcription" (2006), "YOU and THOU in Early Modern English: cross-linguistic perspectives" (2007) and "Handwriting and the linguistic study of letters" (forthcoming).

Roberta Facchinetti, Full Professor of English, *Facoltà di Lettere e Filosofia*, University of Verona. Her main fields of research are textual analysis and the pragmatics of discourse with particular reference to verbal modality through the use of language corpora, both from a synchronic and a diachronic perspective. Among her most recent publications: *Theoretical Description and Practical Applications of Linguistic Corpora* (2007), *English Modality in Perspective. Genre Analysis and Contrastive Studies* (co-edited with Frank Palmer, 2004), *Corpus-based Studies of Diachronic English* (co-edited with Matti Rissanen, 2006) and *Corpus Linguistics 25 Years on* (edited, 2007). Recently she has also dealt with ELT, by co-authoring the *Grammar Units of the Cambridge-CRUI B1 Online Course* (with Sharon Hartle, 2007).

Valerio Fissore, Full Professor of English Language and Translation, *Facoltà di Lingue e Letterature Straniere*, University of Turin. His research is in the areas of geographical varieties of English (Welsh English and Welsh writing in English) and of text types, with a view to complementing a linguistic theory of translation. His recent publications include: “Biforcazioni” (2004), on the strategies of translation, “La discrezione dei fiori” (2005), criticism of translation, “Notes towards a linguistics of verse translation” (2006), and a new translation of T.S. Eliot’s *Four Quartets* (2007), with an introduction to verse translation.

Cristiano Furiassi, Researcher in English Language and Linguistics, *Facoltà di Lingue e Letterature Straniere*, University of Turin. His research interests lie in the field of the lexical contact between English and Italian from a metalexicographic and corpus linguistic perspective. He has published articles on anglicisms and false anglicisms in Italian, discourse analysis and translation, among which “False Anglicisms in Italian monolingual dictionaries: a case study of some electronic editions” (2003), “Spoken interaction and discourse markers in a corpus of learner English” (with Virginia Pulcini, 2004), “Translating American and British trademarks into Italian. Are bilingual dictionaries an aid to the user?” (2006), and “The retrieval of false anglicisms in newspaper texts” (with Knut Hofland, 2007).

Maurizio Gotti, Full Professor of English Language and Translation, *Facoltà di Lingue e Letterature Straniere*, University of Bergamo. He is Director of CERLIS (*Centro di Ricerca sui Linguaggi Specialistici*), the research centre on specialized languages based at the University of Bergamo. His main research area concerns the features and origins of specialized discourse. He is a member of the Editorial Board of national and international journals, and edits the *Linguistic Insights* series for Peter Lang. Among his most recent publications: *Investigating Specialized Discourse* (2005), *Explorations in Specialized Genres* (edited with Vijay K. Bhatia, 2006), *Insights into Specialized Translation* (edited with Susan Sarcevic, 2006) and *Intercultural Aspects of Specialized Communication* (with Christopher N. Candlin, 2007).

Ruth Anne Henderson, Lecturer in English Language and Translation, University of Turin. She is a history of the English language scholar. She has done research in Shakespearian and Renaissance grammar. She is now working on the language of liturgy in relation to culture and language-driven translation strategies. Her recent publications include: “Negative declarative *do* in Shakespeare's prose” (2006), “Muriel Spark. Dio, il diavolo e il doppio” (2007), “Approaches to exposition in Shakespeare’s plays” (2007).

Sara Laviosa, Researcher in English Language and Translation, *Facoltà di Lingue e Letterature Straniere*, University of Bari. Her field of research is translation studies and has published essays and articles in the field of the theory and the teaching of translation. She has edited the first volume of *Translation Studies Abstracts* (1998) and *L'approche Basé sur le Corpus/ The Corpus-based Approach* (1998). She is the author of *Corpus-based Translation Studies. Theory, Findings, Applications* (2002) and *Linking Wor(l)ds. Lexis and Grammar for Translation* (2005).

Aurelia Martelli, Researcher in English Language, *Facoltà di Lingue e Letterature Straniere*, University of Turin. She has participated in the ICLE project (International Corpus of Learner English). Among her publications: "Lexical errors in the semantic field of 'work' in ICLE-IT" (2004), "Cleft sentences in ICLE-IT" (2004) and a monograph on error analysis and learner corpora (*Lexical Collocations in Learner English: A Corpus-based Approach,* 2007).

Gerardo Mazzaferro, Researcher in English Language and Linguistics, *Facoltà di Lingue e Letterature Straniere*, University of Turin. His research has focussed on English historical linguistics and English sociolinguistics. His recent publications include *The English Language and Power* (ed. 2002), *A Short History of English* (2006) and *Global English. Identity, Culture and Power* (forthcoming).

Davide Mazzi, Lecturer in English Language and Translation, *Facoltà di Lettere e Filosofia*, University of Modena and Reggio Emilia. His research interests have focussed on discourse analysis, corpus linguistics, argumentation studies, with particular reference to legal discourse and academic discourse. His recent publications include "The construction of argumentation in judicial texts: combining a genre and a corpus perspective" (2007), "Reporting Verbs: a Tool for a Polyphonic Reading of Judgments" (2007) and *The Linguistic Study of Judicial Argumentation: Theoretical Perspectives, Analytical Insights* (2007).

Vincenza Minutella, Researcher in English Language and Translation, *Facoltà di Lingue e Letterature Straniere,* University of Turin. She has carried out research on theatre translation, film adaptations of literary texts and Shakespeare translation, focussing in particular on 'Romeo and Juliet' and its Italian translations for page, stage and screen, and on audiovisual and film translation. Her publications include "Translating Romeo and Juliet. The sonnet moves back to Italy" (2002) and *Translating for dubbing from English into Italian* (2007).

John Morley, Full Professor of English Linguistics, *Facoltà di Scienze Politiche*, University of Siena. His main research interests are related to corpus linguistics and the media. He is in charge of the Media Working Group of the European project on Citizenship entitled IntUne (Integrated and United: A Quest for Citizenship in an ever closer Europe). He has contributed to national and international journals, see especially "Lexical cohesion and rhetorical structure" (2006) and edited volumes, among which *Massed Medias: Linguistic Tools for Interpreting Media Discourse* (with Linda Lombardo, Louann Haarman, and Christopher Taylor, 1999), *Corpora and Discourse* (with Alan Partington and Louann Haarman, 2004). He is the author of *Truth to Tell: Form and Function in Headlines* (1998).

Andrea Nava, EFL instructor (*collaboratore ed esperto linguistico*), *Facoltà di Lettere e Filosofia*, University of Milan. He holds a postgraduate degree from the universities of Edinburgh, Lancaster and Milan. His main research interests lie in the areas of grammaticography, pedagogical grammar, corpus linguistics and the history of language teaching.

Stefania Nuccorini, Full Professor of English Language and Translation, *Facoltà di Lettere*, University of Roma 3. Her research interests include phraseology, contrastive analysis (English and Italian), lexicology, corpus linguistics, dictionaries and dictionary use. Among her publications: *Phrases and Phraseology: Data and Descriptions* (edited, 2002), "In search of phraseologies: discovering divergences in the use of English and Italian true friends" (2006), "Italian phraseology" (2007), "Because change happenz. On the phraseological environment of 'change' and 'happen'" (2007) and "Note su alcune fraseologie nei dizionari pedagogici più recenti" (2007).
Web page: http://host.uniroma3.it/dipartimenti/linguistica/doc_nuccorini.html

Alan Partington, Associate Professor of English Language, *Facoltà di Scienze Politiche*, University of Bologna. He has published works in the fields of phonetics, CALL, lexicology, discourse analysis, corpus linguistics and the philosophy of language. He is currently researching ways in which corpus techniques can be used to study features of discourse (Corpus-assisted Discourse Studies or CADS). He is the author of *Patterns and Meanings* (1998), *The Linguistics of Political Argument* (2003), "'Utterly content in each other's company': semantic preference and semantic prosody" (2004) and *The Linguistics of Laughter* (2006).
Web page: http://didattica.spbo.unibo.it/pais/partington/index.html

Maria Pavesi, Full Professor of English Language and Linguistics, *Facoltà di Lettere e Filosofia*, University of Pavia. She has been involved in different areas of research including second language acquisition, ESP, corpus linguistics and translation for dubbing. She has taken part in various national and international projects and was one of the main organisers of "Cinema paradiso: i sottotitoli nell'apprendimento linguistico", a project on subtitling in language learning co-financed by the EU in the European year of languages. She has published articles on word formation, linguistic education, film translation, dubbing and subtitling. She is the author of *Formazione di parole. La conversione in inglese L2* (1994), *La traduzione filmica. Aspetti del parlato doppiato dall'inglese all'italiano* (2005), "Lingua di arrivo e lingua di partenza nel doppiaggio dei pronomi" (2008), and "Profiling film translators in Italy: a preliminary analysis" (with Elisa Perego 2006).

Luciana Pedrazzini, Researcher in English Language and Translation, *Facoltà di Lettere e Filosofia*, University of Milan. Her main research interests include second language acquisition, assessment, autonomous learning and learner and teacher corpora. She has published articles on language teaching ("I percorsi di apprendimento autonomo di lingua inglese: un'esperienza graduale di self-access in ambito universitario", 2004) and teacher education ("Come si diventa insegnante ricercatore?", 2003).

Virginia Pulcini, Associate Professor of English Language and Translation, *Facoltà di Lingue e Letterature Straniere*, University of Turin. She has done research on English phonetics and phonology (*Introduzione alla pronuncia inglese*, 1990), lexicography (*La lessicografia bilingue tra presente e avvenire*, edited with Elena Ferrario 2002), spoken discourse ("A corpus of 'informal academic interviews': the Italian Component of the LINDSEI project", 2004), and anglicisms in Italian ("A new dictionary of Italian Anglicisms: the aid of corpora", 2006). She is the Italian coordinator and compiler of the LINDSEI corpus (Louvain International Database of Spoken English Interlanguage). She is currently compiling a dictionary of Anglicisms.

Maria Cecilia Rizzardi, Associate Professor of English Language and Translation, *Facoltà di Lettere e Filosofia*, University of Milan. Her main fields of study are English linguistics, foreign language pedagogy and distance learning. She has published several books and articles on action research, teaching methods, testing and teacher education. She is the author of *Insegnare la lingua straniera. Apprendimento e ricerca* (1997), *Programmare e insegnare le lingue straniere nella scuola di base* (2000) and *Metodi in classe per insegnare la lingua straniera* (with Monica Barsi, 2005).

Martin Solly, Associate Professor of English Language and Linguistics, *Facoltà di Scienze della Formazione*, University of Florence. His main research interests and academic publications concern language learning in higher education and specialized discourse, especially the language of the law. He is currently investigating the lexico-grammatical and textual features of the discourse of insurance, including healthcare insurance, and that of educational reform. His recent publications include "'Don't get caught out': pragmatic and discourse features of informative and promotional texts in international healthcare insurance" (2007) and "Linguistic choice in the discourse of contemporary educational reform" (2007).

Margherita Ulrych, Full Professor of English Linguistics and Translation, *Facoltà di Scienze Linguistiche e Letterature Straniere*, Catholic University of Milan. Her research interests include the integration of descriptive and applied approaches in the theory and practice of translation, the use of computerized corpora in English and translation studies, film translation and English for specific purposes. She has published widely in these fields and her leading works are *Tradurre. Un approccio multidisciplinare* (1997), *Focus on the Translator in a Multidisciplinary Perspective* (1999), *Terminologia della Traduzione* (2002) and *Mediating Discourse. A Descriptive Approach to Applied Translation Studies* (forthcoming). She is editor of the series *Traduzione. Testi e Strumenti* and *TRANSIT. Translation, Mediation.*

English Library:
the Linguistics Bookshelf

Printed Edition – ISSN 1974-0603
Electronic Edition – ISSN 1974-0670

The book series "English Library: the Linguistics Bookshelf" is meant to be a forum for scientific discussion and debate over any topic of English linguistics, in a theoretical, descriptive or applied perspective, both synchronically and diachronically. It aims to provide new insights into English phonetics and phonology, morphology, syntax, lexis, semantics and pragmatics, and the interface between different levels of linguistic analysis. New and recent research methodologies, critical approaches, and specialized fields of knowledge – such as corpus linguistics, critical discourse analysis, translation studies, the varieties of English and ESP – will be dealt with in the series.

“English Library: the Linguistics Bookshelf” addresses a readership composed of academics and students interested in the nature, history and usage of the English language.
The book series will publish monographs, collections of essays and conference proceedings. All the books in the series will be peer-reviewed by an international advisory board co-ordinated by the scientific committee. It is hoped that the online open-access publication of the “English Library: the Linguistics Bookshelf” series will encourage scientific dialogue among researchers in Italy and worldwide.

Publications:

Aurelia Martelli and Virginia Pulcini, eds (2008), *Investigating English with Corpora. Studies in Honour of Maria Teresa Prat*, Polimetrica Publisher, Italy. ISBN 978-88-7699-103-5

Elisa Mattiello (2008), *An Introduction to English Slang. A Description of its Morphology, Semantics and Sociology*, Polimetrica Publisher, Italy. ISBN 978-88-7699-113-4 (forthcoming)

Related series:

Lexicography worldwide:
theoretical, descriptive and applied perspectives
series editor Giovanni Iamartino

Félix San Vicente, ed. (2006), *Lessicografia bilingue e traduzione. Metodi, strumenti, approcci attuali*, Polimetrica Publisher, Italy. ISBN 978-88-7699-048-9

Félix San Vicente, ed. (2007), *Perfiles para la historia y crítica de la lexicografía bilingüe del español*, Polimetrica Publisher, Italy. ISBN 978-88-7699-075-5

Maria Colombo et Monica Barsi, textes réunis par (2008), *Lexicographie et lexicologie historiques du français - Bilan et perspectives*, Polimetrica Publisher, Italy. ISBN 978-88-7699-084-7

Giovanni Iamartino and Nicholas Brownlees, eds (2008), *Insights into English and Germanic lexicology and lexicography: past and present perspectives*, Polimetrica Publisher, Italy. ISBN 978-88-7699-082-3 (forthcoming)