CREATIVE FORUM

Creative Forum – CF, (started in 1988) is a peer-reviewed international journal, which is published in March and September every year. It publishes original research papers pertaining to the late 20th century and current literary practices in India and elsewhere. To outline a few major areas, on which we have centered our previous researches in this journal and intent to invite papers for future issues are:

- Contemporary literature
- Literary theories
- Popular fiction
- Cultural studies
- Feminist studies
- Indian English writing
- Women's writing
- Literary criticism
- Marxist criticism
- Postcolonial studies
- Feminist criticism
- Modernism & postmodernism
- Stylistics
- Structuralism & post-structuralism
- Identity
- Aesthetics

Over a number of years special theme issues devoted to poetry and fiction such as *Recent Indian English Poets*; *Quest for Identity in Indian English Fiction and Poetry*; *Crisis of Identity*; *Commonwealth Literature*, *Post-Colonial Indian English Literature*; *New Zealand Literature*; *Black Literature*; *Stream of Consciousness in Indian English Fiction*; *Indian Literatures in Translation & Partition re/visioned*, *Perspectives on Women's Writing in India*, *Popular Culture Studies*, *Comparative Poetics*, *Dalit Writings & Fiction to Film* etc. have been published.

NOTE FOR THOSE WHO WISH TO CONTRIBUTE

Authors can submit their manuscripts on the above related subjects to the Editors on their e-mail: <bahrius@vsnl.com> in a WORD file according to the camera ready format given on our website: <http://www.bahripublications.in>.

INTERNATIONAL JOURNAL OF TRANSLATION
VOL. 22, NO. 1-2, JAN-DEC 2010

European Parliament Interpreting Corpus (EPIC): Methodological Issues and Preliminary Results on Lexical Patterns in Simultaneous Interpreting

ANNALISA SANDRELLI
Libera Università San Pio V (LUSPIO), Italy

CLAUDIO BENDAZZOLI
MARIACHIARA RUSSO
University of Bologna at Forlì, Italy

ABSTRACT

The European Parliament Interpreting Corpus (EPIC) is one of the first machine-readable corpora available in the field of Interpreting Studies. It was created in 2004/2006 by the Directionality Research Group, based at the University of Bologna at Forlì (Italy), and consists of 9 sub-corpora in total: three sub-corpora of source language speeches (Italian, English and Spanish) and six sub-corpora of simultaneously interpreted speeches, thus comprising all possible directions and combinations of the three languages involved. The present paper focuses on two main areas of Corpus-based Interpreting Studies: methodology and applied research. The first part addresses some of the main methodological issues that arise when creating a machine-readable corpus of simultaneous interpreting (SI) material, particularly in data collection and corpus design. The second part presents the main results of one of the studies carried out on EPIC material so far, namely a study of lexical patterns that draws on Laviosa's study on lexical density and lexical variety in source and target texts of English narrative prose (Laviosa 1998b). The same methodology is applied to all EPIC material, analysed from both a comparable and a parallel perspective. The results thus obtained shed light on the role played by translation mode (written translation vs. simultaneous interpreting), language combination and language direction.

Keywords: Corpus-based interpreting studies, corpus design, parallel corpus, comparable corpus, lexical density, lexical variety.

INTRODUCTION

"Many of the observations encountered in the literature on interpreting are based on sparse, often anecdotal data [...]. The compilation of bilingual and parallel corpora is indeed overdue, given the potential to use large, machine-readable corpora to arrive at global inferences about the interpreted text" (Shlesinger 1998: 487). A decade has gone by, but Shlesinger's statement still holds true. The corpus-based approach[1] in Interpreting Studies (IS) is still far from being as developed and productive as is the case in Translation Studies (TS). In the latter, the cross-fertilisation between corpus linguistics and descriptive studies has led to the creation of many parallel and comparable corpora over the years, with a large amount of authentic data in machine-readable form. These corpora have been exploited for a variety of purposes, which range from the study of translators' language and translation strategies to the testing of translation hypotheses (Baker 1998) and trainee translators' education (among others, Aston 2001; Bernardini 2004). In the field of IS, efforts have been made to expand existing collections of data from professional and educational settings in order to study interpreting norms and linguistic patterns, but as yet they have not produced large machine-readable corpora.[2] This possibility is mentioned by Setton (2001) in his paper on the analysis of interpreting corpora. Recently, an interesting attempt was made by Shlesinger (2009), who compiled a machine-readable comparable corpus (English and Hebrew) to study the difference in the linguistic richness and lexico-grammatical features in the outputs of six professional interpreter-translators. The corpus contains the transcripts of six interpretations and six translations of the same source text. It is the first comparable intermodal corpus to provide evidence of the peculiarities of two different language-transfer modalities, i.e. interpreting and written translation, produced by the same subjects.

The stark contrast between TS and IS in their approach to corpus linguistics, and the consequent lack of interpreting corpora available to the scientific community, can be partially explained by the many hurdles intrinsic to research on conference interpreting and to spoken corpora in general. These difficulties are analysed in greater detail in this paper (see Section 1.1). Here, it will merely be mentioned that the extremely time-consuming task of transcribing spoken data and the misconceived idea that a corpus-based approach entails merely quantitative analysis are among the factors that may have discouraged the exploring of this virtually untrodden path.

To rise to this challenge of compiling an interpreting corpus, an interdisciplinary research group was set up at the University of Bologna (Forlì), comprising interpreter trainers and researchers, computational linguists, corpus linguists and IT technicians[3] to create the European Parliament Interpreting Corpus (EPIC).[4] The interplay of individual members' expertise, supported by adequate funding which allowed three of its members to work full-time on the project (one member for one year and two for two years), was a *conditio sine qua non* for the implementation of a project of this scope, the primary objective of which was to study the effect of directionality and language-pair specificities at all levels (lexical, morphosyntactical and pragmatic) by means of corpus linguistics tools and semi-automated processes. Given its enormous potential as a scientific and pedagogical resource, EPIC is at present being accessed for teaching purposes, graduation theses and scholarly research.

The research presented in this paper concerns two specific features of interpreters' output compared with the output of source language speakers, namely lexical density and lexical variety. The study of lexical density in speech is a promising approach to assess information density in interpreting, since verbal information is mostly conveyed through content – or lexical – words (nouns, verbs, adjectives and adverbs) rather than function words (prepositions, determiners, pronouns, conjunctions, particles, numerals, interjections, negatives, etc.). Furthermore, according to Halliday (1993, as quoted in Castello 2004), lexical density is a distinctive measure of the difference between spoken and written language since it is likely to be twice as high in the written mode. In addition, the lexical variety parameter provides an insight into an interpreter's expressive skills and linguistic richness. Our study was inspired by Laviosa's paper (1998b), in which she compared lexical density and lexical variety in English narrative prose and in English translations from various languages.

In this paper, the relationship between corpus linguistics and simultaneous interpreting (SI) is discussed in Section 1, with special reference to the obstacles encountered in compiling a corpus (Section 1.1) and to the development stages of the EPIC project (Section 1.2); section Section 2 contains a description of EPIC and its search tools; section Section 3 describes the methodology adopted to study lexical density and lexical variety in the corpus (Section 3.1), analyses the results obtained in the English sub-corpora of EPIC (source and target speeches) and compares them with Laviosa's findings (Section 3.2); there follows an overview of lexical density and lexical variety patterns

across all of EPIC's nine sub-corpora in Section 3.3 and Section 3.4 respectively. Finally, tentative conclusions are put forward as regards this complex scenario in Section 4.

1. CREATING SI CORPORA

Looking at the Interpreting Studies literature, the term "corpus" is probably used as frequently as in research and publications on written translation. However, there is less availability of *machine-readable* interpreting corpora (Pöchhacker 2002: 104), i.e. material that can be studied using corpus linguistics tools and made available to researchers to replicate previous studies. Indeed, many researchers use relatively small collections of transcripts of source and target speeches; moreover, the relevant analogue recordings or digital audio files are seldom available to the research community as a whole[5]. In other words, occurrences of specific patterns are generally counted manually on the basis of *ad hoc* annotations developed by individual researchers (Setton, forthcoming).

The fact that so few interpreting researchers have adopted the corpus-based approach can be partially explained by the many obstacles and the great number of issues related to corpus design and corpus compilation (Armstrong 1997). In addition to all the general methodological issues involved in spoken corpora development (Thompson 2005), the creation of interpreting corpora raises even more complex issues, which depend on the interpreting mode concerned. For example, building a corpus to study and/or teach consecutive interpreting implies a number of methodological choices which are different, to a certain extent, from those involved in the creation of a corpus of material for simultaneous interpreting, whispered interpreting, liaison interpreting, sign language interpreting, and so on. These differences are due not only to the intrinsic features of each interpreting mode *per se* (consider, for example, recording equipment or transcription visualisation options), but also to the communicative situations in which these services are provided. Moreover, even after data collection has been successfully completed, the next stage in corpus compilation is not free of obstacles (see Section 1.1 below).

With regard to the various interpreting modes mentioned above, here the focus is on simultaneous interpreting. In particular, we describe some of the challenges researchers are faced with when compiling an SI corpus. It must be borne in mind, however, that most of the considerations also apply to other interpreting modes. Then, we explain how these challenges were addressed in the development of the EPIC

corpus, in an attempt to share a number of suggestions with readers who may wish to use them in their own research endeavours.

1.1. *Methodological hurdles*

Over the last few decades simultaneous interpreting has become one of the most widespread forms of interpreting (Riccardi 2003: 106). Since its early days after World War I and its "official debut" after World War II (Baigorri 2000), its use has been expanding from international organisations to local markets in several countries, covering a virtually unlimited number of sectors and subjects. The greater availability of sources of SI material is both an opportunity and a challenge for researchers. It is an opportunity in that interpreting is a very specific activity that is more circumscribed than other forms of human communication (Gile 2001: 7). Its growth has meant larger markets and, consequently, a higher number of professional interpreters, who now constitute a much larger population than fifty years ago, with an increasing proportion of women (Baigorri 2003). On the other hand, it is also a challenge as the many diverse contexts mean not only different communication patterns, but also different degrees of accessibility and cooperation from potential providers of material (i.e. interpreters, speakers, technical staff, organisers, etc.).

Existing sources of SI material include international settings, such as the institutions of the European Union and the United Nations (Baigorri 2004). In some bilingual countries (e.g. Canada), national parliaments carry out their activities with a permanent team of simultaneous interpreters. Universities can also be sources of SI data: interpreter training schools are ideal places to collect material to study trainees' performance in a learner corpus (Lindquist 2005); moreover, academic SI is routinely offered by professional interpreters in universities in multilingual societies, as is the case of South Africa (van Rooy 2005; Wallmach 2006). Finally, local markets in several countries are "teeming with" SI interpreter-mediated events, such as court proceedings, conferences, business meetings, TV and radio shows, film festivals and so on.

Such a great variety of sources poses a number of common challenges to researchers, who need to deploy different strategies to address them effectively (Gile 1998). Challenges range from sample representativeness (made difficult by the high number of variables), to availability and accessibility of material (i.e. collaboration from interested parties and consent), data collection practices, transcription, annotation and encoding standards, and data accessibility and exchange

(Cencini 2002). Of course, many of these obstacles also affect empirical research in interpreting in general, even when it is carried out using more "traditional" approaches. However, working with corpora presupposes collecting large quantities of data, which makes these obstacles and challenges even more critical.

The following subsection briefly describes how these difficulties were tackled in the creation of the European Parliament Interpreting Corpus (for a more detailed account, see Monti, Bendazzoli, Sandrelli & Russo 2005; and Bendazzoli & Sandrelli 2005).

1.2. *Compiling the EPIC corpus*

The use of material from EU institutions, in our case the European Parliament, to study SI provides several advantages.[6] One of them is that there are more constant variables than in material taken from any local market-based setting, thus ensuring good levels of representativeness (Halverson 1998). EP plenary sittings are held regularly throughout the year and this is set to continue in the future. Furthermore, the debates are structured and organised according to established rules and procedures; participants (e.g. MEPs, Commissioners, and so on) have a precise political role and may take the floor to perform specific speech events chosen from a narrow range of options (e.g. presenting a report, opening a session, asking a question, raising a point of order, etc.). Topics discussed at the EP vary enormously, but the communicative goals are the same in every sitting and what is debated is usually the result of discussions held in parliamentary committee meetings over the previous weeks or months. Interpreters are all professionals and they are carefully selected by means of specific accreditation tests. They normally work into their mother tongue, though after the latest enlargement rounds *retour* (i.e. working into one's foreign language) is being used more commonly for less available language combinations (Marzocchi & Zucchetto 1997).

Another advantage is that accessibility to this material is less problematic than it can be in many other settings. EP plenary sittings are broadcast live by *Europe by Satellite* (EbS). This satellite TV channel allows users to select the language channel from among all the EU official languages, so it is possible to record the source language speakers from the floor and interpreters from their booths by using different TV sets. In this respect, it must be pointed out that recently both accessibility and ease of use have been greatly improved, as it is now possible to download digital video files of the full sittings from the EP website (as of April 2006). All the EPIC material comes from a

number of European Parliament plenary sittings held in 2004[7]; EbS broadcast only parts of the debates, so our recordings do not include the full plenary sittings. All the material was recorded on videotapes, which were subsequently digitised and edited by using dedicated software; the resulting digital files were stored in a multimedia archive (the EPIC archive).

Is it possible, however, to freely use such an abundance of SI material for academic purposes? Researchers are well aware of the fact that they are likely to come across copyright limitations when studying real life material. Indeed, if "it is harder and less straightforward to obtain permission for translated text" (Baker 1996: 178), it can be all the more challenging to obtain permission for interpreted text. Fortunately, since all activities carried out within EU institutions must comply with the principle of transparency, permission is granted to anyone who may want to use recordings of EP debates and SI for academic (non-commercial) purposes. Moreover, the EP Audiovisual Services can be contacted for help in obtaining specific material.

Another unique characteristic of the EP as a source of SI material is that there are twenty-three official languages currently in use. In compliance with Rule 138, "All Members shall have the right to speak in Parliament in the official language of their choice" (European Parliament 2006: 65-66): this makes it possible to obtain and analyse multiple target versions into different languages of the same source speeches (Vuorikoski 2004).

Finally, it is worth highlighting that background material is also available to describe EP communicative activities in more detail. The EP website contains abundant documentation and information about MEPs, procedures and activities, not to mention the verbatim reports drafted by EP officials for each plenary sitting. These reports are a faithful written rendition of what is said in every debate and they also include the official headings of the agenda. The verbatim reports are usually drafted first in the original language(s) used during the debates; then this version is translated into all the other EU official languages. In our case, since our intention was to collect and classify data in three languages only (out of the then 20 official languages), the information in the verbatim reports was used to identify the participants who used Italian, English or Spanish in the debate, and to select the relevant parts of the digital recordings available in our EPIC multimedia archive. Unsurprisingly, the written excerpts found in the verbatim reports could not be used as "transcripts" because they did not fully reflect the wording of each speech, nor did they include spoken language features (e.g. repetitions, false starts, unfinished words, mispronunciations, and

so on). Furthermore, this method to select relevant speeches implied that some speech events were excluded outright, because they are not usually recorded in the EP verbatim reports (e.g. the President giving the floor to speakers and most communicative exchanges during voting time).

Once the data collection process has been completed, researchers are only at the beginning of their work. To start with, (video or audio) recordings must be transcribed for later analysis. Transcription is an extremely time-consuming activity, which can quickly lead researchers to reshape the size and scope of their studies, even after obtaining large quantities of genuine recordings (Kalina 1994). In the EPIC project, teamwork and the use of speech recognition software considerably helped in speeding up this labour-intensive but essential stage (the transcribers involved are all trained conference interpreters and were able to perform shadowing to transcribe interpreters' output. In other words, transcribers listened to target speech recordings and simultaneously repeated them aloud – obviously, the speech recognition software had been previously trained to recognise their voice).

Transcription conventions are likely to vary from study to study, on account of two interrelated aspects: on the one hand, researchers decide what and how to transcribe depending on their specific research questions and objectives; on the other hand, representing spoken language features in writing is not a straightforward task, especially when one attempts to produce transcripts that are both user-friendly and machine-readable (Chafe 2005; Cook 2005; Shlesinger 1998). Machine-readable transcripts are one of the fundamental ingredients of electronic corpora, but encoding schemes (i.e. the way structural and linguistic information is described in a corpus) are also needed in order to organise material in a structured way and to be able to retrieve occurrences automatically. Once again, there are currently very few examples of shared sets of attributes (e.g. nonverbal features one might want to include and annotate, or established classifications of speech events, participants, etc.) and encoding schemes – one exception is the suggested application of the Text Encoding Initiative scheme to the Television Interpreting Corpus, in Cencini (2002) and Cencini & Aston (2002). The lack of a common standard is a major problem when it comes to comparing results, replicating studies and merging different sets of transcripts to create larger corpora. For this reason, particular attention should be paid to saving material in widely compatible formats (e.g. .txt for text files and .wav for audio files), so as to be able to run them on different software programmes without problems. In our case, we opted for an orthographic transcription with only a limited

number of paralinguistic features (see Monti et al. 2005). Furthermore, each EPIC transcript includes a header with extra-linguistic information about the speech and the speaker (see Section 2.1). For the time being, EPIC transcripts are accessible online (see Note 5), but the multimedia archive is only accessible on the local area network (LAN) of our Department.

A further step in the creation of an interpreting corpus is alignment. There can be two levels of alignment. Firstly, each transcript can be aligned with its recording (for example, by using a software programme called *Transana*, see Web references). This is the best option to carry out studies on spoken corpora, given the fact that transcripts are only a partial representation of the actual data under investigation, i.e. the audio/video recordings. Secondly, interpreting corpora can also be aligned by matching source texts (STs) with their target texts (TTs). In the case of SI corpora, this second type of alignment can be carried out not only on the basis of content, but also on the basis of the real-time delivery of source and target speeches, i.e. considering the time span between the speaker's and the interpreter's outputs. This time-based alignment seems possible if dedicated software is used, such as *Winpitch* or *Exmaralda* (see Web references). At present, EPIC has not been aligned in any of the above-mentioned ways, but it has been fully POS-tagged, lemmatised and indexed (see Section 2.3). Therefore, it is possible to carry out studies on the frequency of words, grammatical structures, discourse patterns, co-occurrences and lexical density (to name but a few research avenues), as advocated by Shlesinger (1998).

2. DESCRIPTION OF THE EPIC CORPUS

The European Parliament Interpreting Corpus is made up of nine sub-corpora in total. It includes three sub-corpora of source speeches in Italian, English and Spanish (indicated as org-it, org-en and org-es) and 6 sub-corpora of (interpreted) target speeches into these three languages, in all possible combinations and directions (indicated as 'int' followed by the language direction, e.g. int-en-it for English into Italian). The material already transcribed and available for analysis is just a portion of the EPIC multimedia archive, corresponding to parts of the European Parliament (EP) plenary sittings held in February 2004. Its current size and composition are displayed in Table 1 below:

Table 1. *EPIC size and composition*

Sub-corpus	No. of Speeches	Total Word Count	% of EPIC
Org-en	81	42,705	25
Org-it	17	6,765	4
Org-es	21	14,406	8
Int-it-en	17	6,708	4
Int-es-en	21	12,995	7
Int-en-it	81	35,765	20
Int-es-it	21	12,833	7
Int-en-es	81	38,066	21
Int-it-es	17	7,052	4
TOTAL	**357**	**177,295**	**100**

2.1. *The EPIC header*

As was mentioned in Section 1.2, each transcript features a header containing information related to participants and speech events. Table 2 presents all the fields related to participants, while Table 3 contains all the information available about each speech event.

Table 2. *Attributes assigned to participants in EPIC*

Attributes	Attribute Allocation Guidelines
Speaker	Surname, first name
Gender	F M
Country	Italy …
Mother tongue	Yes No
Political function	MEP MEP Chairman of the session President of the European Parliament Vice-President of the European Parliament European Commission European Council Guest
Political group	Verts/ALE PPE-DE PSE ELDR GUE/NGL UEN TDI EDD NI

Table 3. *Attributes assigned to speech events in EPIC*

Attributes	Attribute Allocation Guidelines
Duration	Short (< 120 seconds) Medium (120 - 360 seconds) Long (long > 360 seconds)
Timing	(Total number of seconds)
Text length	Short (< 300 words) Medium (300 - 1000 words) Long (long > 1000 words)
Number of words	(Total number of words)
Speed	Slow (<100 w/m) Medium (100 - 120 w/m) High (> 120 w/m)
Words per minute	(Number of words per minute)
Source text delivery	Read Impromptu Mixed
Topic	Agriculture & Fisheries Economics & Finance Employment Environment Health Justice Politics Procedure & Formalities Society & Culture Science & Technology Transport
Specific topic	(as indicated in the verbatim report)

Interestingly, the values we had assigned to some speech event-related attributes had to be re-adjusted within a certain range to fit the specificity of the material included in EPIC. More specifically, although "duration" and "speech length" were classified as short, medium or long, whereas "speed of delivery" (number of words per minute) as low, medium or high, the actual ranges indicated in Table 3 can only be considered valid within the context of EP debates, during which 150 w/m can be considered a "medium" speed of delivery. Mode of delivery was also classified by using three categories, depending on whether speakers could be seen reading a script (read mode), or speaking (seemingly) without the aid of any written material (impromptu delivery), or switching between reading and speaking off-the-cuff (mixed mode).

Finally, the last header attribute was reserved for comments (Table 4), so as to be able to include a variety of details, which were deemed useful, but which would not fit in the previous categories.

Table 4. *Attributes assigned to "comments" in EPIC*

Attributes	Attribute Allocation Guidelines
Comments	(specify Council configuration)
	(specify Commission DG)
	(specify title) e.g. President of the Republic of Colombia
	(specify accents) e.g. Scottish accent, Irish accent
	Technical problems

2.2 *Speeches in EPIC*

2.2.1. *English source speeches and corresponding target texts*

The sub-corpus named *org-en* includes source speeches delivered in English and is the largest in EPIC, accounting for almost 24% of the overall word count. It comprises 81 speeches, 3 of which delivered by non-native speakers (from Denmark, the Netherlands and Portugal). Thirty-five speeches are delivered by Irish speakers and 43 by British speakers. The majority of speakers are men (65 vs. only 16 women). As can be expected, most speeches are delivered by Members of the European Parliament (56), but there are also some speeches made by European Commissioners (18) and Ministers of the European Council (7).

Turning to the characteristics of the English source speeches, more than half are read from a written script (43 out of 81), whereas just over one fourth (24) are delivered impromptu. The remaining speeches (14) are delivered in a mixed mode, i.e. switching between reading and improvising. In terms of duration, half of the speeches are medium (40) by EP standards, lasting between 2 and 6 minutes. Twenty-eight speeches are short, and only 13 are classified as long. Thus, average duration is around 3 minutes and 30 seconds.[8] Clearly, text length (i.e. word count) reflects similar patterns, in that over half of the English source speeches (44) are of medium length, 27 speeches are short and only 10 speeches were long.[9] With regard to speed of delivery, interestingly, there are almost as many speeches delivered at a fast pace (34) as at a medium pace (36). The average speed across the org-en sub-corpus is 156.5 w/m. Finally, the topics discussed in these speeches are varied, ranging from politics to health to economics. This also applies to the other sub-corpora of source speeches; however, in the case of English source speeches there is a large number of speeches belonging to the Procedures and Formalities category, which is related to formulaic language used in the European Parliament (e.g. by the

Chairman of a debate). This is because Pat Cox was President of the EP and Ireland held the EU presidency at that time.

Let us now describe the main features of the output of EP interpreters who dealt with these source texts in their booths. We shall look at the Italian booth first and then the Spanish one. The *int-en-it* sub-corpus is the largest collection of Italian target speeches. In the vast majority of speeches it is possible to hear female interpreters (68 occurrences vs. 13 male voices). The average speed of delivery is 123.7 w/m per minute (much lower than that of the English source speeches). Similarly, the *int-en-es* sub-corpus is the largest of the three Spanish sub-corpora in EPIC. More than half of its 81 speeches are delivered by female interpreters (58 occurrences vs. 23 speeches with voices of male interpreters). Interpreters work at a medium speed of about 137 words per minute (slower than the English source speakers, but faster than the Italian interpreters). For both sets of target speeches, although duration is basically the same as the source speech (owing to the simultaneity of the translation process), text length is shorter than the source speeches, i.e. the EP interpreters in the Italian and Spanish booths produced a lower number of words than the English source speakers (Italian booth – 6,940 words; Spanish booth – 4,639 words).

2.2.2. *Italian source speeches and corresponding target texts*

The *org-it* sub-corpus comprises 17 Italian source speeches delivered by native Italian speakers. They are all MEPs – 14 men and 3 women – belonging to different political groups. This is the smallest sub-corpus of source language speeches, comprising, only 6,765 words in total (see Table 1). Eight speeches are read out written texts, 6 are delivered off-the-cuff and 3 are delivered in a mixed mode. In terms of duration, 13 speeches are classified as medium and only 4 as short (there are no long speeches). The overall duration of Italian source speeches amounts to almost 50 minutes, with an average duration of 3 minutes per speech. There are 10 medium-length and 7 short speeches, with an average count of about 400 words per speech. Unexpectedly, speed of delivery is low in 11 speeches and medium in 6 speeches. On average, this set of Italian speeches is delivered at a speed of about 130 words per minute – by EP standards, this seems to be comfortable for interpreters translating from Italian.

Let us turn to the target speeches into English and into Spanish. The *int-it-en* sub-corpus is the smallest one in EPIC: it comprises 17 target speeches delivered by male interpreters in 8 cases and female interpreters in 9 cases; one interpreter has a non-native accent. As regards speed, the average value is 132.2 w/m, i.e. slightly faster than

the SL speeches average. The *int-it-es* sub-corpus is the smallest of the three Spanish sub-corpora. Of the 17 speeches, in ten cases the interpreter's voice is female, while the remaining seven speeches are interpreted by male interpreters. The average speed here is about 136 words per minute (note that the average speed of the Italian SL speakers was around 130 words per minute).

As was the case for the previous group of target speeches, the duration of source and target speeches is the same, but in terms of text length there is a slight drop (-57 words) in the word count of the English target speeches (int-it-en) and an increase (by almost 300 words) in the Spanish target speeches (int-it-es). This might be an effect of language-pair and language direction, in that interpreters were working from a Romance language into a Germanic language (Italian into English) in one case and from a Romance language into another Romance language (Italian into Spanish) in the other.

2.2.3. *Spanish source speeches and corresponding target texts*

The *org-es* sub-corpus contains the source speeches delivered in Spanish. It is double the size of its Italian counterpart, but it is considerably smaller than the sub-corpus of English source speeches. It accounts for about 8% of the entire EPIC corpus and includes 21 speeches, with an overall duration of nearly two hours. The speeches are delivered by 14 male speakers and 7 female speakers. The majority of speeches are delivered by MEPs, followed by speeches by European Commission representatives. In addition, there is one speaker who does not have any political role within the EU institutions and who is not Spanish: Colombia's President Uribe, who addressed the House on 10 February 2004. It is interesting to note that the speech delivered by the only non-EU guest is the longest in terms of duration and number of words.

Turning to the mode of delivery, most speeches are either read from written scripts (9 instances) or delivered in a mixed mode (7 instances). Only 5 speeches are delivered entirely off-the-cuff. In terms of temporal duration, only 4 speeches are classified as long (i.e. over 6 minutes), 12 speeches fall into the medium category and 5 are short. This is reflected in the text length parameter, i.e. the number of words in each speech: there are 4 long speeches, 10 medium ones and 7 short ones. Finally, the data on speed (number of words per minute) show that most speeches are delivered at a very fast pace: in ten instances at more than 160 words per minute, seven speeches are delivered at medium speed, while only four are delivered at low speed, with an overall average of about 152 w/m.

Let us now consider the target speeches into English and into Spanish. In the *int-es-en* corpus, there are 16 speeches interpreted by male interpreters and 5 by female interpreters, all of them native speakers. Their average speed of delivery is 136.2 w/m. In the *int-es-it* sub-corpus there are only female interpreters and their average speed of delivery is 124.5 words per minute. As in the previously analysed sub-corpora (see Section 2.2.1 and Section 2.2.2), both sets of target speeches are characterised by a lower number of words than their Spanish source speeches (-1,411 words in the English booth and -1,573 in the Italian booth). This seems to be a general trend across the whole corpus, as the only exception was found in the int-it-es sub-corpus.

2.3. *EPIC research tools and methods*

As was briefly mentioned at the end of Section 1.2, the EPIC corpus is fully POS-tagged, lemmatised and indexed. Part-of-speech tagging can be done automatically by using dedicated software programmes called taggers. The main stages of the tagging process are tokenisation, tag assignment and disambiguation (Bowker & Pearson 2002). Tokenisation means breaking down the text into individual tokens, i.e. words and punctuation signs, if there are any (in EPIC there is no punctuation since it is typical of the written mode and this is a spoken corpus). Then, the tagger assigns a part-of-speech (POS) tag to each item on the basis of morphological and context-based cues. Different taggers use different methods to do this.

The taggers chosen for EPIC, namely *Treetagger* (Schmid 1994) for English, *Freeling* (Carreras, Chao, Padró & Padró 2004) for Spanish and the combination of taggers suggested by Baroni, Bernardini, Comastri, Piccioni, Volpi, Aston & Mazzoleni (2004) for Italian,[10] are all stochastic taggers, which means that they "generally resolve ambiguities by using a training corpus to compute the probability of a given word having a given tag in a given context" (Jurafsky & Martin 2000: 300).[11]

The taggers chosen for EPIC have very high accuracy rates; it must be noted, however, that they were developed for written texts, not for transcripts of spoken texts. We expected them to run into difficulties when encountering certain features of our corpus that are typical of spoken language and of the EP setting in particular. Therefore, we carried out a study to calculate the accuracy rate for all three taggers by selecting a random sample from all nine sub-corpora and manually checking both tagging and lemmatisation results in each of them (Sandrelli & Bendazzoli 2006). The results were very encouraging since

Treetagger achieved an average accuracy rate of 97% across the three English sub-corpora, an average of 92% was obtained in the Italian sub-corpora, while *Freeling* obtained an average of about 95% in the Spanish sub-corpora.

Nevertheless, certain items in our transcripts were found to be problematic for all three taggers. Firstly, specific tags will have to be added to the taggers' current tagsets to enable them to deal with filled pauses (*ehm*) and truncated words, both of which are typical features of the EPIC transcripts. Moreover, specific rules will have to be devised to enable the taggers to distinguish capitalised nouns, referring to people, places, institutions, legislation, etc., which are very frequent in our corpus. Lists of items (nouns, verbs, or adjectives) and interjections (*please*, *thank you*, *hello*, etc.) also misled the taggers in some cases, owing to the absence of any punctuation-related information. A form of prosodic annotation (with prosodic breaks inserted in appropriate places) may be a suitable solution to this problem.

The above-mentioned problems are examples of aspects that must be taken into account when deciding whether to tag a corpus and when selecting the best tagger with which to do it. Although the tagging process is automatic, improving a tagger's performance and making results more reliable may be a time-consuming step.

The tagged transcripts were also encoded by means of the *IMS Corpus Work Bench – CWB* (Christ 1994), which associates positional attributes to all individual words in the corpus and XML structural attributes to the header fields in the transcripts (see Tables 2, 3 and 4 in Section 2.1). This means that it is possible to query the corpus in the CQP language of *CWB*, either via the Unix command line or through a dedicated web interface available on the Forlì School for Translators and Interpreters' development website (see Web references), to retrieve and analyse material of interest. An example of the tagged and encoded corpus can be seen below, in which the XML attributes are followed by a column containing the tokens, a second column with the POS-tags, a third column of lemmas, and a fourth column containing a transcription of how the word was actually uttered. Clearly, in most cases column 1 and column 4 are identical, but when there is a disfluency, column 1 contains the "correct" form of the word and the last column an orthographic transcription of how the word was actually pronounced (see below *stupplying* instead of *supplying*):

```
<speech date="10-02-04-m" id="005" lang="en" type="org-
en" duration="long" timing="392" textlength="medium"
length="906" speed="medium" wordsperminute="139"
delivery="read" speaker="Byrne, David" gender="M"
country="Ireland" mothertongue="yes" function="European
Commission" politicalgroup="NA" gentopic="Health"
sptopic="Asian bird flu" comments="Health and Consumer
protection; Irish accent">
I          PP     I          I
have       VHP    have       have
been       VBN    be         been
supplying  VVG    supply     /stupplying/
[...]
</speech>
```


2.3.1. *Search tools: The web interface*

As well as allowing access to the EPIC corpus, the web interface also features a number of information pages on the EPIC project, including the EPIC multimedia archive, transcription criteria, EP debates, and so on. Potential users are required to register on the web page: registration is granted by the site administrator to anyone who may wish to use or take a look at EPIC for research purposes.

The web interface enables users to carry out both simple and advanced queries, and to produce word frequency lists. However, since EPIC is made up of nine sub-corpora, the first step is to select one of them and then query it. In other words, since EPIC has not been aligned, if the aim is to compare source texts with their corresponding TTs, separate queries must be made in the relevant corpora. Similarly, if one wants to use EPIC as a comparable corpus rather than as a parallel one, e.g. if the objective is to study certain characteristics of speeches originally produced in English and then compare them with speeches interpreted into English, the English ST sub-corpus (org-en) and the English TTs sub-corpora (int-it-en & int-es-en) must be queried separately.

After selecting the desired sub-corpus, if users choose the simple query option, they can either search through the whole sub-corpus or restrict the search to a number of texts by using one or more of the search filters provided. These are based on the attributes used to classify speakers and speeches in the header of each transcript (see Section 2.1). For example, users may want to find all occurrences of a certain word or phrase only in Economics speeches, or only in speeches delivered by a man, or only in long speeches, and so on.

The EPIC web interface also enables users to issue advanced queries formulated in the CQP language of CWB.[12] Users can search the corpus by POS-tag(s) or lemma, or by combining a word search and a POS-tag search: for example, all the instances of the English auxiliary

"to be" followed by an "-ing" form can be retrieved automatically and compared with their Italian and Spanish renditions in the corpus.

The results of both simple and advanced queries are visualised in a KWIC (key-word-in-context) view, with the queried string in the middle and the specified left and right contexts. The full text in which a specific occurrence was found can also be displayed in order to look at a larger context. It is also possible to display the XML attributes containing speaker and speech information, the 'normalised' text with no disfluencies (option "Show Word"), the transcript with any disfluencies orthographically transcribed ("Show Transcript"), the lemmatised transcript ("Show Lem"), or the transcript with the part-of-speech tags ("Show Pos").

2.3.2. *Search methods: Data extraction examples*

The lengthy process required to compile EPIC (see §1.2 and 2.3.1) had the objective of creating an electronic collection of data that could be searched automatically. An example of a study in which data extraction was made easier by the fact that EPIC is a machine-readable corpus is our study of disfluencies in interpreting (Bendazzoli, Russo & Sandrelli forthcoming). The aim was to investigate the possible correlation between two types of disfluencies – mispronounced words and unfinished words – in STs and TTs, i.e. to see whether language production errors made by source language speakers induced similar errors in the interpreters' output or whether interpreters produce their own disfluencies independently. The transcription conventions adopted in EPIC (see Monti et al. 2005) were exploited to find all the occurrences of these two types of disfluencies. More specifically, in EPIC truncated words are conventionally transcribed with a dash where the truncation occurs (e.g. *it is an interesting* ***pro-*** *proposal*). This feature was exploited to extract all instances of truncated words from the corpus, basically instructing the computer to retrieve all instances of words ending in a dash.

As regards mispronounced words, the process was marginally more complex. As was explained in Section 2.3.1, the tokenised and tagged files of the corpus feature four columns for each token, and the fourth column contains the orthographic transcription of disfluencies. Therefore, a command was issued asking the computer to search for all those cases in which there is a difference between column 1 and column 2 of each tokenised file. The same procedure was repeated for all the 9 sub-corpora, making it possible to obtain the full list of such disfluencies in EPIC in a matter of minutes.

Another example of how the characteristics of EPIC have made it easier to obtain interesting data for analysis is our study on lexical density and lexical variety (Sandrelli & Bendazzoli 2005; Russo, Bendazzoli & Sandrelli 2006), which is discussed more in detail in Section 3. As was mentioned in the Introduction, the study was inspired by an article on lexical density in translational vs. non-translational texts (Laviosa 1998b), which concluded that lexical density is lower in translated texts than in narrative prose originally written in English. Our interest lay in verifying whether similar conclusions could be reached about interpreted speeches vs. speeches originally delivered in the same language.

Lexical density is expressed as the number of lexical words divided by the overall number of running words in a corpus (Stubbs 1986, 1996; Laviosa 1998b). In order to calculate lexical density in our sub-corpora, we therefore had to count all lexical (content) and function words in each sub-corpus. Separate lists of content words and function words were extracted automatically thanks to the POS tags assigned to each word in EPIC. In order to create the list of lexical words in the org-en sub-corpus, the computer was asked to retrieve all the words whose POS was noun, adjective, verb or adverb; as regards function words, they include prepositions, determiners, pronouns, conjunctions, particles, numerals, interjections, negatives, greetings, and politeness markers. The two lists obtained were then manually checked and "cleaned," i.e. words mistakenly assigned to the wrong list because of incorrect tagging were moved to the correct one.[13] This procedure was followed for each sub-corpus under analysis, and lexical density was calculated for each of them. Unlike the previous example on disfluencies, in this case data extraction was automatic, but also required extensive revision work. It was undoubtedly faster and more accurate, however, than it would have been if it had been done manually.

Laviosa also found that high frequency words are repeated more often in translated texts than in original prose (see Section 3.1 for more details). In testing her conclusion against our material, we were able to exploit the fact that EPIC has been tagged and encoded, and is therefore very easy to use to produce word frequency lists.

The above examples demonstrate the usefulness of the laborious compilation process of our corpus: the two studies referred to in this paragraph (on disfluencies and on lexical density and variety in interpreting) would have been very complex and, above all, time-consuming to carry out if all the relevant data had been selected manually. The following paragraph gives a fuller account of the study

on lexical density and variety by providing further information on its methodology and discussing its results.

3. Lexical Density and Lexical Variety

3.1. *Definitions and methodology*

Our research took as a starting point a study by Laviosa on lexical density and variety in translational vs. non-translational texts (Laviosa 1998b). Her corpus of translated texts included narrative prose from the Translational English Corpus (TEC). Laviosa found four main lexical patterns in TEC, three of which were tested against our material too, namely (Laviosa 1998b: 563):

i) Translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower);
ii) The proportion of high frequency words versus low frequency words is relatively higher in translated texts;
iii) 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).[14]

In our first paper on lexical patterns (Sandrelli & Bendazzoli 2005), we only looked at the English and Italian parts of EPIC, including both original and interpreted speeches. Data on lexical patterns in the Spanish sub-corpora in EPIC (org-es, int-en-es and int-it-es) were added in a later study (Russo et al. 2006), in which we also drew some conclusions on lexical patterns in interpreted speeches in general, thereby shedding some light on the role played by language pair and language direction.

It is worth pointing out that Laviosa's original and translated texts were all in English, whereas we analysed Italian, English and Spanish sub-corpora (separately). In order to compare our corpus with hers, the same methodology and the same definitions were used in our study. As regards calculating lexical density, for example, we applied the following definition: "Lexical density is expressed as a percentage and is calculated by subtracting the number of function words in a text from the number of running words (which gives the number of lexical words) and then dividing the result by the number of running words" Laviosa (1998b: 565).

In order to verify whether Laviosa's conclusions on lexical variety (her second and third finding on the proportion of high frequency vs. low frequency words) also applied to EPIC, word frequency lists were produced for our nine sub-corpora and then the top 100 words were selected from each of them (list heads). Then, the overall word count

was obtained for each list head, and the percentage of sub-corpus covered by each list head was calculated. In other words, if the 100 most frequent words in a sub-corpus account for a large part of that sub-corpus, lexical variety is low, because it means that the same words are used over and over again.

3.2. *Results: A comparison between Laviosa's findings in TEC and our findings in EPIC*

In order to ascertain whether Laviosa's conclusions on lexical density and lexical variety in English translated texts also apply to interpreted output, we shall take a look at the three English sub-corpora in EPIC first, i.e. the sub-corpus of SL English speeches (org-en) and the sub-corpora of interpreted speeches from Spanish and Italian into English (int-es-en and int-it-en), which make up most of EPIC (see Table 1 and Section 2.2.1). Afterwards, we extended our analysis to the other languages in our corpus to detect any trends emerging from the whole of EPIC (see Section 3.3. and Section 3.4).

Before presenting the results of our study, a few remarks need to be made on our respective corpora. There is a considerable difference in size between the two corpora studied by Laviosa, i.e. the Non-Translational English Corpus (NON-TEC) and the Translational English Corpus (TEC), and the corresponding EPIC sub-corpora of English original speeches and English interpreted speeches (the latter include the speeches interpreted from both Italian and Spanish into English, see Table 5). One must not forget, however, that Laviosa's material consists of written texts, whereas EPIC contains transcripts of speeches delivered orally and interpreted simultaneously. Therefore, even though a corpus of as many as 3 million words does not seem large enough to corroborate any hypotheses by corpus linguistics standards (Baker 1998: 482), for an electronic corpus of its kind, the size of EPIC is nonetheless considerable. As more transcripts are added over time, hopefully our corpus-based data will become more and more indicative of general trends.

Table 5. *Total running words and lexical density in NON-TEC, TEC and EPIC (English sub-corpora only).*

Corpus	Total Word Count	Lexical Density
NON-TEC	729,349	54.9536606
TEC	999,945	52.87439153
EPIC-org-en	42,705	57.31179018
EPIC-int-en (int-it-en + int-es-en)	19,703	57.30599401

Let us now turn to the results emerging from a direct comparison of Laviosa's data with our English SL speeches and English interpreted outputs, starting with lexical density (see Table 5). The effect noted by Laviosa (1998b) in translated texts, i.e. a lower lexical density than in texts originally written in English (52.874 vs. 54.954), was not confirmed by our study. In fact, there is a negligible variation in lexical density when the totality of interpreted speeches is compared with original English speeches (57.306 vs. 57.312).

The first observation about this general finding is that lexical density in these spoken sub-corpora is remarkably higher than in Laviosa's written corpora. This seems to counter previous distinctions between oral and written modalities: Ure (1971, as quoted in Castello 2004) found that written texts tend to have higher lexical density (over 40%) and, as already mentioned, Halliday (1993, as quoted in Castello 2004) stated that lexical density in written texts is likely to be twice as high as in oral texts because of nominalisation processes. Stubbs (1996, in Castello 2004: 133) calculated lexical density in written and spoken texts, and found that the former range between 40% and 65% and the latter from 34% to 58%. Our results betray the "hybrid" nature of our texts, a majority of which are written to be spoken (classified as "read") or partly planned on paper ("mixed"), with a small minority of impromptu speeches, which are never fully improvised (see Section 2.2.1), owing to the institutional setting and the communicative situation (see Section 2.2). In an oral-to-written continuum, these speeches are closer to the written end. Secondly, the virtually overlapping results obtained for original and interpreted speeches seem to confirm the suggestion made by Shlesinger (1989, 2009) "whereby interpreting exerts a leveling effect: oral texts become more literate, literate texts become more oral."

However, a language-pair specific difference emerges if the two EPIC sub-corpora of interpreted speeches are compared separately (see Table 6 below). In the sub-corpus of speeches interpreted from Spanish into English, lexical density is slightly lower (-0.22060126), whereas in the speeches interpreted from Italian into English it is actually slightly higher than in the original English speeches (+0.41033265).

Table 6. *Lexical density in each of the EPIC English sub-corpora*

Sub-corpus	Total Running Words	Lexical Words	Function Words	Lexical Density
org-en	42,705	24,475	18,230	57.31179018
int-it-en	6,708	3,872	2,836	57.72212283
int-es-en	12,995	7,419	5,576	57.09118891

This suggests that a role is played by the source language (SL): we could hypothesise that the Italian SL speeches contained a larger proportion of lexical words than the Spanish SL speeches, and that this may have influenced the linguistic choices made by the English interpreters to convey the speaker's message. Indeed, the overall picture of our results on lexical density across EPIC (see Table 8 and §3.3) indicates a higher lexical density in the Italian SL speeches than in the Spanish ones. So, the SL factor associated to target text (TT) lexical density might be worth considering. However, a word of caution is necessary since our corpus of Italian original speeches is at present quite small and further data are called for before we can draw any conclusions.

The second objective of the present paper was to verify Laviosa's findings on lexical variety, i.e. that translated texts feature a higher proportion of high frequency words versus low frequency words. As was explained in detail in Section 3.1, the first 100 words in our frequency lists were selected to create our list heads. Then the total word count of each list head was considered in order to calculate the percentage of each list head in the respective sub-corpus. We also looked at the percentage of lexical vs. function words in the list heads. This was not calculated by Laviosa, so no comparison is possible. The data obtained for the English sub-corpora compared with Laviosa's results are presented below (Table 7).

Table 7. *Lexical variety in Laviosa's corpora and in the EPIC English sub-corpora.*

Sub-corpus	List Head Word Count	% of Sub-corpus	Lexical Words in List Head		Function Words in List Head	
			Word Count	% of List Head	Word Count	% of List Head
org-en	**22,745**	**53.26**	6,142	27.00	16,603	73.00
int-it-en	**3,832**	**57.09**	1,250	32.60	2,582	67.40
int-es-en	**7,176**	**55.22**	2,112	29.40	5,064	70.60
NON-TEC	**380,226**	**51.60**				
TEC	**537,900**	**56.20**				

The data in Table 7 are in line with Laviosa's findings for translational English. The percentage of high frequency words in the list heads is higher for interpreted English than original English by a considerable margin (+ 3.83% for *int-it-en* and + 1.96% for *int-es-en*). These data indicate that the nuclei of words most frequently used in speeches

interpreted into English are less varied and account for a larger part of the corresponding sub-corpora. Given such a finding, the linguistic output of interpreters might possibly be perceived as being less refined in terms of style, considering that linguistic variety is generally one of its constituents.

A similar result regarding lower lexical variety in oral interpretations versus written translations of the same source text was achieved by Shlesinger (2009), who applied a different method – the ratio of types to tokens – to calculate linguistic richness in her intermodal corpus (see Introduction): the type-token ratio of the spoken corpus was 0.655, while that of its written counterpart was 0.735. She carried out both an intra-subject and inter-subject evaluation, and so the only variable determining this finding seems to be the translational activity.

As for the distribution of lexical and function words in the EPIC list heads, the English source speeches show a lower percentage of lexical words than both interpreted English sub-corpora. This may indicate the tendency of interpreters to reformulate their output (by adding synonyms or explanations), to insert self-corrections, or to expand and explain the source text, which would make interpreted texts richer in lexical words (and possibly more informative) than speeches originally produced in English.

In the list heads of each English sub-corpus, function words are generally twice as numerous as content words (see percentages in Table 7). As already mentioned, function words include articles, pronouns, most prepositions, conjunctions, some adverbs, auxiliary verbs (be, have, do) and modal verbs. As Shlesinger (2009) noted "they are often associated with the explicitness and redundancy of the spoken language." A corollary of the above observation about the difference in the percentage of lexical words between the list head of our English sub-corpora is the fact that here function words occur more often in source speeches (73.00%) than in speeches interpreted from Italian (67.40%) and from Spanish (70.60%). Since source speeches in EPIC are mainly read speeches, as was observed earlier, we would have expected to find more function words in interpreted speeches, which are "truly" spoken outputs. This finding may be due to a variety of reasons: corpus size effect (the interpreted sub-corpora are considerably smaller), the above-mentioned features typical of interpreting, better text-condensing strategies by interpreters, or the already mentioned interpreting levelling effect that makes oral speeches more literate and vice versa (Shlesinger 1989, 2009). In order to test these hypotheses

concerning the role of lexical and function words in interpreted speeches, it would be useful to have the interpreted speeches aligned with their source speeches, which is one of our intended future developments.

3.3. *Lexical density: Trends and patterns across EPIC*

After comparing Laviosa's data with the results obtained for the English part of our corpus (the only ones that could be directly compared), in this paragraph the full set of data on lexical density in EPIC is presented and commented. Given the very specific structure of EPIC, it is possible to look at it both in a comparable corpus perspective, i.e. by comparing lexical density across the sub-corpora grouped by language (the three English sub-corpora, the three Italian ones and the three Spanish ones) and in a parallel corpus perspective, i.e. comparisons are made between speeches originally delivered in one language and their corresponding two interpreted versions.

Let us start with the comparable perspective that also inspired Laviosa's work. As can be seen from Table 8 below, the trend observed is the opposite of that noted by Laviosa in relation to translated texts, i.e. lexical density tends to be higher in interpreted speeches than in speeches originally delivered in the same language. There are only two exceptions, namely the Spanish into English sub-corpus (int-es-en), in which lexical density is slightly lower than in the sub-corpus of speeches originally delivered in English (org-en); and, above all, the Spanish into Italian sub-corpus, in which lexical density is decidedly lower than in speeches originally delivered in Italian (almost -1%).

Table 8. *Lexical density across EPIC: comparable perspective*

Sub-corpus	Total Running Words	Lexical Words	Function Words	Lexical Density
org-en	42,705	24,475	18,230	**57.31179018**
int-it-en	6,708	3,872	2,836	**57.72212283**
int-es-en	12,995	7,419	5,576	**57.09118891**
org-it	6,765	3,997	2,768	**59.08351811**
int-en-it	35,765	21,209	14,556	**59.30099259**
int-es-it	12,833	7,452	5,381	**58.06904075**
org-es	14,406	7,762	6,644	**53.88032764**
int-it-es	7,052	3,836	3,216	**54.39591605**
int-en-es	38,066	20,702	17,364	**54.38449009**

It is impossible to say, at this stage, whether it is a coincidence or not that the two exceptions to the trend concern speeches interpreted from the same source speeches (org-es), i.e. whether lower lexical density in

these two cases may be related in some way to the specific composition of the org-es sub-corpus. However, it must be highlighted that the sub-corpus of speeches originally delivered in Spanish (org-es) does have some unique features. The group of Spanish source speakers is the only one to include a non-EU guest, the President of Colombia, whose speech actually makes up roughly one fourth of the entire sub-corpus, with a duration of about thirty minutes (see Section 2.2.3).

The same results can be analysed from an interlingual perspective, i.e. using EPIC as a parallel corpus. Since lexical density in interpreted speeches is compared with lexical density in their source speeches, the same results on lexical density are grouped differently in Table 9 for the sake of clarity: each sub-corpus of source speeches is followed by the corresponding sub-corpora of interpreted speeches.

Table 9. *Lexical density across EPIC: parallel perspective*

Sub-corpus	Total Running Words	Lexical Words	Function Words	Lexical Density
org-en	42,705	24,475	18,230	57.31179018
int-en-it	35,765	21,209	14,556	59.30099259
int-en-es	38,066	20,702	17,364	54.38449009
org-it	6,765	3,997	2,768	59.08351811
int-it-en	6,708	3,872	2,836	57.72212283
int-it-es	7,052	3,836	3,216	54.39591605
org-es	14,406	7,762	6,644	53.88032764
int-es-en	12,995	7,419	5,576	57.09118891
int-es-it	12,833	7,452	5,381	58.06904075

The first aspect that can be noted is that all the corpora of interpreted speeches are smaller than their respective corpora of source speeches, i.e. the total number of running words in interpreted sub-corpora is lower than the number of running words in the original sub-corpora. This seems to indicate a certain amount of text compression in all language directions, with the exception of the Italian into Spanish sub-corpus (+287 words). In this particular case, the higher number of running words is to be attributed exclusively to the increased presence of function words (+448) compared with the corpus of Italian source speeches. However, it is interesting to see that the only other case of an increase in the number of function words in the whole of EPIC is the int-it-en sub-corpus, i.e. the corpus of speeches interpreted into English from the same source speeches (org-it). It seems that both the Spanish and the English interpreters at work on those particular Italian speeches (org-it) were forced to make use of more function words than usual, i.e.

more function words than those used in the Italian source speeches and proportionally more function words than those used when interpreting from other languages (compare with the other data in the table). This could be related to the characteristics of the Italian source speeches in question. For example, it could be hypothesised that in these Italian speeches links were not always explicit enough, or that the average length and complexity of Italian sentences forced interpreters to break them down into more manageable units and to add conjunctions. A detailed analysis of these Italian source speeches is needed before a satisfactory explanation can be found.

3.4. *Lexical variety: Trends and patterns across EPIC*

As illustrated earlier (Section 3.1), the second objective of our study was to verify Laviosa's findings on lexical variety, i.e. that translated texts reveal a higher proportion of high frequency words versus low frequency words. The results obtained on lexical variety in the whole of EPIC are discussed here following the same order of presentation as in the previous section on lexical density (Section 3.3), i.e. by considering EPIC first as a comparable and then as a parallel corpus.

Looking at lexical variety in the EPIC list heads from a comparable perspective (Table 10), we obtained the following results:

Table 10. *Comparable analysis of EPIC list heads.*

Sub-corpus	List Head Word Count	% of Sub-corpus	Lexical Words in List Head		Function Words in List Head	
			Word Count	% of List Head	Word Count	% of List Head
org-en	22,745	**53.26**	6,142	27.00	16,603	73.00
int-it-en	3,832	**57.09**	1,250	32.60	2,582	67.40
int-es-en	7,176	**55.22**	2,112	29.40	5,064	70.60
org-it	3,365	**49.74**	892	26.50	2,473	73.50
int-en-it	17,353	**48.51**	4,771	27.50	12,582	72.50
int-es-it	6,264	**48.82**	1,572	25.10	4,692	74.90
org-es	7,825	**54.30**	1,719	22.00	6,106	78.00
int-it-es	4,021	**57.00**	1,033	25.70	2,988	74.30
int-en-es	21,087	**55.40**	4,922	23.34	16,165	76.66

As can be seen in the third column of percentages, the results obtained for the English list heads are similar to the results obtained for the Spanish list heads, i.e. speeches interpreted into these two languages feature a higher percentage of high frequency words than source speeches in the same languages (in other words, the TL list heads

account for a larger area of each respective sub-corpus). This means that both interpreted English and interpreted Spanish in EPIC have a lower lexical variety than original English and original Spanish (in line with Laviosa's findings on translational English). By contrast, the opposite trend was found in the Italian material. Here, the list heads of target speeches from both English and Spanish source speeches (int-en-it & int-es-it) cover a smaller area of each respective sub-corpus and their percentages of high frequency words are lower than Italian source speeches (org-it). In other words, original Italian appears to have a lower degree of lexical variety than interpreted Italian. This may suggest greater attention to lexical variety on the part of Italian interpreters, irrespective of language-pair specificity.

As for the distribution of lexical vs. function words in the list heads, there seems to be a higher percentage of lexical words (and, inversely, a lower percentage of function words) in interpreted speeches than in original speeches of the same language. This trend is particularly marked in the Italian into Spanish and the Italian into English direction, whereas it is less so in the English into Spanish and the Spanish into English direction. This may reflect a corpus size effect, as the sub-corpus of Italian source speeches is much smaller than the Spanish and English ones, or an influence of lexical patterns in the Italian source speeches (as suggested earlier, Section 3.2).

Let us now consider the data on list heads from a parallel perspective, i.e. comparing results obtained from source speeches with the results obtained from their corresponding target speeches. Table 11 below presents the overall picture:

Table 11. *Parallel analysis of EPIC list heads*

Sub-corpus	List Head Word Count	% of Sub-corpus	Lexical Words in List Head		Function Words in List Head	
			Word Count	% of List Head	Word Count	% of List Head
org-en	22,745	**53.26**	6,142	27.00	16,603	73.00
int-en-it	17,353	**48.51**	4,771	27.50	12,582	72.50
int-en-es	21,087	**55.40**	4,922	23.34	16,165	76.66
org-it	3,365	**49.74**	892	26.50	2,473	73.50
int-it-en	3,832	**57.09**	1,250	32.60	2,582	67.40
int-it-es	4,021	**57.00**	1,033	25.70	2,988	74.30
org-es	7,825	**54.30**	1,719	22.00	6,106	78.00
int-es-en	7,176	**55.22**	2,112	29.40	5,064	70.60
int-es-it	6,264	**48.82**	1,572	25.10	4,692	74.90

Starting with the two sub-corpora of target speeches from English (int-en-it & int-en-es), the percentage of high frequency words in the int-en-it sub-corpus is markedly lower than the percentage in the org-en sub-corpus; the opposite is true of target speeches into Spanish. This means that lexical variety in Italian interpreted speeches is higher than in Spanish interpreted speeches from the same source language. If we look at the internal distribution of lexical vs. function words in these list heads, there is almost no variation in Italian interpreted speeches in comparison with the English source speeches. The only difference concerns the number of both lexical and function words (Italian target speeches feature a considerable reduction). The situation is different in the int-en-es sub-corpus, in which the percentage of function words is higher than in the English source speeches (76.60% vs. 73.00%), meaning that there is more widespread use of high frequency function words.

In the sub-corpora of target speeches interpreted from Italian (int-it-en & int-it-es), the percentages of high frequency words are higher than in the source Italian speeches (org-it), and they are also very similar to the percentages found in the English and Spanish target versions (respectively, 57.09% and 57.00%). This means that, in this case, the two sub-corpora of interpreted speeches display less lexical variety than the sub-corpus of Italian source speeches. However, the internal distribution of lexical and function words in the list heads is different in the two target languages: the English list head is characterised by greater use of high frequency lexical words, whereas the Spanish one is characterised by greater use of high frequency function words than their corresponding source material.

Finally, lexical variety appears to be lower in the Spanish into English direction and higher in the Spanish into Italian direction compared with the percentage of high frequency words in the *org-es* list head. Looking at the internal composition of this last group of list heads, both sub-corpora of target speeches present a higher percentage of lexical words than the source speeches, with a considerable increase in the Spanish into English direction.

Overall, what can be seen by analysing the list heads in EPIC is that lexical variety in the TTs is generally lower than in their STs . This trend is particularly marked in the sub-corpora of target speeches interpreted from Italian as there was a considerable increase in the percentage of high frequency words in both sub-corpora. The only exceptions to this pattern are the speeches interpreted into Italian from both English and Spanish (int-en-it & int-es-it) as the list heads for

these two directions account for a lower percentage of sub-corpus than their corresponding source speeches. This seems to indicate more lexical variety in the output of Italian interpreters than in the source speeches delivered by English and Spanish language speakers in EPIC.

4. CONCLUSIONS

The present paper examines some of the many methodological issues in Corpus-based Interpreting Studies and provides an example of applied research, presenting a study on lexical density and lexical variety in the European Parliament Interpreting Corpus (EPIC).

The creation and development of machine-readable corpora is less straightforward in Interpreting Studies than it is in Translation Studies. This is mostly due to the greater number of methodological hurdles researchers have to overcome when compiling an interpreting corpus. These range from the high number of variables involved in data collection, to corpus design, transcription, annotation standards, and data accessibility and exchange.

EPIC is one of the first attempts to overcome such obstacles. Despite its current limited size (approximately 177,000 tokens overall) and partial accessibility (transcripts can be queried through a web-based interface, but video and audio digital files are only accessible on the LAN of our Department), it is a first step towards providing the research community with a freely accessible electronic tool to be exploited for both research and training purposes. It includes three sub-corpora of source speeches (in Italian, English and Spanish) and six sub-corpora of target speeches (covering all possible combinations and directions between the three languages concerned). The corpus does not feature any type of alignment yet (neither text-to-sound, nor text-to-text, be it content-based or time-based). However, its structure makes it possible to carry out research from multiple perspectives, i.e. EPIC can be used as both a comparable and parallel corpus. Furthermore, a future application to be considered is using EPIC as an intermodal corpus (Shlesinger 2009), studying EP source speeches, target speeches and the written translations published in the verbatim reports.

The European Parliament as a communicative framework and its plenary sittings as a communicative situation proved to be a fruitful source of SI material for several reasons: it ensures high levels of representativeness in terms of participants, text types, speech events, interpreters' levels of expertise, working conditions, and so on; material can be easily accessed and permission to use it for academic purposes

can be obtained; plentiful background material is available on the EP website and in relevant publications (e.g. the verbatim report and its translated versions); it is a unique communicative context, in which the same source speech is now interpreted into 22 different languages during plenary part-sessions.

In the second part of the paper, we presented a study on lexical density and lexical variety in EPIC, whose results are compared with those obtained by Laviosa (1998b) in a study on translational and non-translational English narrative prose. Our results on lexical density go against Laviosa's conclusions on translated texts, as lexical density tends to be higher in the English interpreted speeches than in English original speeches in EPIC. Indeed, this is a prevailing trend across the whole of EPIC.

The higher lexical density found in interpreted speeches seems to blur the distinction between oral and written texts based on this parameter (Ure 1971; Halliday & Martin 1993), although Stubbs (1996) indicated less clear-cut thresholds. Our data seem to support the suggestion that interpreting tends to make oral texts more literate and literate texts more oral (Shlesinger 1989; 2009). Furthermore, higher lexical density could be seen as an indication of interpreters' more detailed presentation of information items, or possibly, as evidence of typical interpreting features, such as repairs, i.e. stylistic or semantic self-corrections and additions of lexical material to make the message more explicit or produce synonymic pairs. These operations are regularly observed in our teaching and professional settings, and are widely documented in the literature (among others, see Petite 2003, 2005; Bendazzoli 2002 on repairs; Schjoldager 1995/2002 & Micheli 2007 on additions). Clearly, these hypotheses on interpreting patterns will have to be tested by carrying out specific studies on the presence and frequency of such verbal material in interpreted speeches.

However, the result obtained is significant in itself, i.e. a difference has emerged concerning lexical density in two types of Translation modes – written translation and simultaneous interpreting. Such a difference is not entirely surprising, given the body of existing literature on cognitive processes and strategies in translation and interpreting (see for example Danks et al. 1997).

It is also worth highlighting an exception to the pattern on lexical density. In the sub-corpus of speeches interpreted from Spanish into Italian, and, to a lesser extent, in the sub-corpus of speeches interpreted from Spanish into English, lexical density was found to be lower than in speeches originally delivered in Spanish. It is impossible, at this stage, to provide a satisfactory explanation for this difference, which does not

appear to be language-pair-related. It would seem to be more closely linked to the nature of the Spanish source speeches in question.

Another tentative conclusion that could be drawn from our study relates to lexical density and the very nature of texts. Ure, Halliday and Stubbs observed that lexical density is related to the mode of expression: written texts are more dense than spoken texts (quoted in Castello 2004). However, comparing Laviosa's results on lexical density in English narrative prose in TEC with our own results on spoken English sub-corpora in EPIC, the pattern suggested by these authors does not seem to be confirmed. Indeed, the percentages of lexical density found in our corpus are higher than those found by Laviosa (around 57% in both original and interpreted English speeches in EPIC vs. 54.95% and 52.87% in original and translated texts in TEC). Interestingly, lexical density patterns in EPIC are consistent within individual languages (about 57% in English, 58-59% in Italian and 53-54% in Spanish). This further confirms that speeches typically interpreted at the EP are closer in nature to written texts along the written-to-oral continuum (source speeches are often read or presented in a mixed mode, mostly at neck-breaking speed): here lies the greatest difficulty for interpreters.

Furthermore, lexical density in speeches interpreted into the same language generally seems to be affected only to a limited extent by language pair and corpus size. For example, interpreting between cognate languages (from Italian into Spanish) or between a Germanic and a Romance language (from English into Spanish) produced similar percentages of lexical density (54.39 and 54.38 respectively), despite the fact that the average speeds of the source speeches differed considerably (130 w/m vs. 152 w/m). Spanish interpreters, irrespective of the language from which they work, seem to display a constant tendency to convey information with the same lexical density. This is further confirmed by the absence of a potential corpus size effect (note that the int-en-es sub-corpus is more than five times as big as the int-it-es one).

The analysis of the entire EPIC corpus has brought to the surface a general tendency towards text compression in interpreted speeches, as indicated by the lower word counts in the sub-corpora of target speeches compared with their source speeches. Once again, there is only one exception to this pattern, this time concerning the Italian into Spanish sub-corpus, which is slightly larger than its source sub-corpus (org-it). We have verified that this increase is caused by a much higher number of function words, in contrast with all the other language combinations and directions. In other words, the general tendency is for

both lexical and function words to decrease in interpreted speeches, resulting in an overall drop in word counts.

However, it should be borne in mind that overall word count is not a meaningful data *per se*. It must be correlated with other information in order to obtain a clearer picture. In our case, we calculated the percentage of high frequency words in each sub-corpus (through the creation of list heads) to obtain an indication of lexical variety. The two are inversely correlated, i.e. a high percentage of high frequency words in a sub-corpus means that a small number of word types (100 in each list head) is used very frequently across the sub-corpus, and therefore that lexical variety is low. The main conclusion that can be drawn from our data in this respect is that the speeches interpreted into Italian during these particular EP sittings displayed more lexical variety than the speeches originally delivered in Italian on the same occasion. This is in contrast with the trend that can be noted for interpreted English and interpreted Spanish, in which lexical variety is lower than original English and original Spanish, in line with Laviosa's conclusions for translated English.

The present investigation into lexical patterns in a trilingual corpus has highlighted interesting trends in the language used by speakers and interpreters within the context of the European Parliament. It is worth stressing that it was only possible to bring these patterns to the surface by means of corpus linguistics tools. The creation of large machine-readable corpora may be seen as a daunting task, in that it involves many theoretical and practical issues, some of which were discussed in this paper. However, this next step in Interpreting Studies seems crucial now if we wish to detect and investigate interpreting norms and test hypotheses on the basis of extensive quantities of genuine data. Such *epic* efforts call for the collaboration of interdisciplinary teams, in which expertise and creativity are pooled, and require substantial funding. Our EPIC experience shows that the manifold research and training applications of these projects make it worth the effort.

NOTES

1. *The Corpus-based Approach* is the title of the insightful special issue of *Meta* edited by Sara Laviosa (1998a). It includes several outstanding contributions, which testify to the degree of theoretical debate and practical applications achieved in Translation Studies almost a decade ago. The contributors tackled many topics, including methodological issues such as adequate representativeness of the data collected in relation to the object of enquiry (Halverson 1998), data-based evidence of translating

patterns such as normalisation (Kenny 1998) and experiences of corpora applications in the training of translators (Zanettin 1998).

2. However, some of the manually collected and analysed corpora are remarkable achievements: two examples are the television interpreting corpus collected by Straniero Sergio (2007), the largest of its kind, including the very first TV appearances of interpreters in Italy, and Vuorikoski's collection of transcribed European Parliament speeches in English, German, Finnish and Swedish (122 source language speeches) and their interpretations into the same four languages (Vuorikoski 2004).
3. Members of the Directionality Research Group: Marco Baroni, Claudio Bendazzoli, Annalisa Sandrelli, Cristina Monti, Gabriele Mack, Elio Ballardini, Peter Mead, Silvia Bernardini and Mariachiara Russo (Project leader).
4. EPIC is an on-line free resource now available at <http://sslmitdev-online.sslmit.unibo.it/index.php>. This Project was made possible thanks to two research grants provided by the Scuola Superiore di Studi Umanistici of the University of Bologna directed by Umberto Eco, and thanks to a research assistant scholarship provided by our Department.
5. Besides this limitation, in IS there are few examples of research based on corpora of transcripts that are also aligned with their respective audio or video files. One such example is the corpus used by Meyer (2008).
6. EP material is also used in SI training (de Manuel Jerez 2003).
7. More specifically, we recorded one part-session and a "shorter" two-day part-session in February, two part-sessions in March, one in April and one in July. The July part-session is already part of the following parliamentary term (2004-2008).
8. These data reflect the specificity of the EP setting, since in most conference interpreting settings speakers tend to take the floor for longer periods.
9. The slight discrepancy between the figures related to duration and text length is caused by the fact that the number of words in a speech depends not only on its duration, but also on the speaker's delivery rate.
10. Fairly obviously, in order to tag texts in different languages, different tagsets and rules must be used because of grammatical differences between languages.
11. In other words, a training corpus is manually annotated and then the tagging algorithm extracts rules from it. When the whole corpus is tagged, the tagger takes decisions on the basis of what it has "learnt" from the training corpus. When a tagger encounters an unknown word, it performs probability calculations and assigns the most likely tag.
12. An information sheet with query suggestions is available on the web site.
13. Certain words proved problematic because of the multiple functions they can have in a sentence (e.g. adjectives or pronouns; adverbs or conjunctions; adverbs or prepositions, etc.). The issue is also discussed in Castello (2004).

14. The fourth finding concerned the number of lemmas in the list heads of translated texts. This particular aspect was not taken into account in our study (see Section 3.1).

REFERENCES

Armstrong, S. 1997. Corpus-based methods for NLP and translation studies. *Interpreting*, 2/1-2, 141-162.

Aston, G. (ed.) 2001. *Learning with Corpora*. Bologna: Clueb.

Baigorri Jalón, J. 2000. *La interpretación de conferencias: el nacimiento de una profesión. De París a Nuremberg*. Granada: Comares.

——. 2003. Conference interpreting: evolution and revolution. Notes on the feminisation of the profession. In M. Z. Gonçalves de Abreu & M. De Castro (Eds.), *Estudos de Traduçao: Actas de Congresso Internacional* (pp. 27-34). Cascais: Principia.

——. 2004. *Interpreters at the United Nations: A History*. Trans. Anne Barr. Salamanca: Ediciones Universidad de Salamanca.

Baker, M. 1996. Corpus-based translation studies: The challenges that lie ahead. In Somers Harold (Ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager* (pp. 175-186). Amsterdam/Philadelphia: John Benjamins.

——. 1998. Réexplorer la langue de la traduction: une approche par corpus. *Meta*, 43/4, 480-485.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. & Mazzoleni, M. 2004. Introducing the La Repubblica corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper in Italian. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa & R. Silva (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (Vol. 5, pp. 1771-1774). Lisbon: ELRA.

Bendazzoli, C. 2002. Repair Strategies and Creativity in Simultaneous Interpretation. Unpublished dissertation, University of Bologna, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Forlì.

——. & Sandrelli, A. 2005. An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In S. Nauert (Ed.), *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation - Saarbrücken 2-6 May 2005*. Available online: <http://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html>.

——., Russo, M. & Sandrelli, A. (forthcoming). Disfluencies in simultaneous interpreting: A corpus-based analysis. *Corpus-based Translation Studies: Research and Applications*. Paper delivered at *Second IATIS Conference, Intervention in Translation, Interpreting and Intercultural Encounters*, University of the Western Cape, South Africa, 12-14 Jul 2006.

Bernardini, S. 2004. Corpus-aided language pedagogy for translator education. In K. Malmkjaer (Ed.) *Translation in Undergraduate Degree Programmes* (pp. 97-112). Amsterdam/Philadelphia: Benjamins.

Bowker, L. & Pearson, J. 2002. *Working with Specialized Language: A Practical Guide to using Corpora*. London/New York: Routledge.

Carreras, X., Chao I., Padró, L. & Padró, M. 2004. Freeling: An open-source suite of language analyzers. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa & R. Silva (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (Vol. 1, pp. 239-242). Lisbon: ELRA.

Castello, E. 2004. Calcolo della densità lessicale e dell'intricatezza grammaticale di corpora linguistici. In C. Taylor Torsello, M. G. Busà, & S. Gessato (Eds.), *Lingua inglese e mediazione interlinguistica. Ricerca e didattica con supporto telematico* (pp. 131-151). Padova: Unipress.

Cencini, M. 2002. On the importance of an encoding standard for corpus-based interpreting studies. Extending the TEI scheme. *Intralinea*, Special Issue CULT2K. Available online: <http://www.intralinea.it/specials/cult2k/ita_more.php?id=107_0_42_0_C>.

——. & Aston, G. 2002. Resurrecting the corp(us|se). Towards and encoding standard for interpreting data. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century. Challenges and Opportunities. Selected Papers from the 1st Forlì Conference on Interpreting Studies, 9-11 November 2000* (pp. 47-62). Amsterdam/Philadelphia: John Benjamins.

Chafe, W. 2005. Adequacy, user-friendliness, and practicality in transcribing. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer* (pp. 54-61). New York: Longman.

Christ, O. 1994. A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94, Budapest*. 12 Mar 2006. Available online: <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/#Papers>.

Cook, G. 2005. Theoretical issues: Transcribing the untranscribable. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer* (pp. 35-53). New York: Longman.

Danks, J., Shreve, G. M., Fountain S. B. & Mc Beath, M. K. (eds.) 1997. *Cognitive Processes in Translation and Interpreting*. ThousandOaks, London and New Delhi: Sage Publications.

de Manuel Jerez, J. 2003. El canal EbS en la mejora de la calidad de la interpretación: perfiles profesionales de especialidad en el itinerario de interpretación. In A. Collados Aís, M. A. Fernández Sánchez & D. Gile (Eds.), *La evaluación de la calidad en interpretación: Investigación* (pp-207-218). Granada: Comares.

European Parliament 2006. *Rules of Procedure of the European Parliament. 16th Edition - July 2006*. Available online: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+RULES-EP+20060703+0+DOC+PDF+V0//EN&>

Gile, D. 1998. Observational studies and experimental studies in the investigation of conference interpreting. *Target*, 10/1, 69-93.

——. 2001. Selecting a topic for PhD research in interpreting. In D. Gile, H. V. Dam, F. Dubslaff, B. Martinsen & A. Schjoldager (Eds.), *Getting Started in Interpreting Research. Methodological Reflections, Personal Accounts and Advice for Beginners* (pp.1-22). Amsterdam/Philadelphia: John Benjamins.

Halliday, M. A. K. & Martin, J. R. 1993. *Writing Science*. London: The Falmer Press.

Halverson, S. 1998. Translation studies and representative corpora: Establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study. *Meta*, 43/4, 494-514.

Jurafsky, D. & Martin J. H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Kalina, S. 1994. Analyzing interpreters' performance: Methods and problems. In C. Dollerup & A. Lindegaard (Eds.), *Teaching Translation and Interpreting 2. Insights, Aims and Visions. Papers from the Second Language International Conference Elsinore, 1993* (pp. 225-232). Amsterdam/Philadelphia: John Benjamins.

Kenny, D. 1998. Creatures of habit? What translators usually do with words. *Meta*, 43/4, 515-523.

Laviosa, S. (ed.) 1998a. *The corpus-based approach, Meta*. 43/4, 474-664. Available online: <http://id.erudit.org/iderudit/003424ar>.

——. 1998b. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43/4, 557-570.

Lindquist, P.P. 2005. Technologies, discourse analysis, and the spoken word: The MRC approach: an empirical approach to interpreter performance evaluation and pedagogy. *Meta*, 50/4. Available online: <http://www.erudit.org/livre/meta/2005/000215co.pdf>.

Marzocchi G. & Zucchetto, G. 1997. Some considerations on interpreting in an institutional context: the case of the European Parliament. *Terminologie et Traduction*, 3, 70-85.

Meyer, B. 2008. Interpreting proper names: Different interventions in simultaneous and consecutive interpreting? *trans-kom, Journal of Translation and Technical Communication Research*, 1/1, 105-122. Available online: <http://www.trans-kom.eu/ihv_01_01_2008.html>.

Micheli, N. 2007. Interpretazione simultanea al Parlamento europeo: il fenomeno delle aggiunte. Unpublished dissertation, University of Bologna, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Forlì.

Monti, C., Bendazzoli, C., Sandrelli, A. & Russo, M. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta*, 50/4. Available online: <http://www.erudit.org/livre/meta/2005/000217co.pdf>.

Petite, C. 2003. Repairs in simultaneous interpreting: Quality improvement or simple error correction?. In Á. Collados Aís, M.ª M. Fernández Sánchez & D. Gile (Eds.), *La evaluación de la calidad en interpretación: investigación* (pp. 61-71). Granada: Comares.

——. 2005. Evidence of repair mechanisms in simultaneous interpreting: A corpus-based analysis. *Interpreting*, 7/1, 27-49.

Pöchhacker, F. 2002. Researching interpreting quality: models and methods. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century. Challenges and Opportunities. Selected Papers from the 1st Forlì*

Conference on Interpreting Studies, 9-11 November 2000 (pp. 95-106). Amsterdam/Philadelphia: John Benjamins.

Riccardi, A. 2003. *Dalla traduzione all'interpretazione. Studi sull'interpretazione simultanea*. Milano: LED.

Russo, M., Bendazzoli, C. & Sandrelli, A. 2006. Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: extended analysis of EPIC (European Parliament Interpreting Corpus). *FORUM, International Journal of Interpretation and Translation*, 4/1, 221-254.

Sandrelli, A. & Bendazzoli, C. 2005. Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, No. 1. Available online: <www.corpus.bham.ac.uk/PCLC>.

——. & Bendazzoli, C. 2006. Tagging a corpus of interpreted speeches: The European Parliament Interpreting Corpus (EPIC). *Proceedings of the LREC 2006 Conference, Genova, Magazzini del Cotone 24-26 May 2006*. Genova: ELRA.

Schjoldager, A. 1995/2002. An exploratory study of translational norms in simultaneous interpreting. Methodological reflections. In F. Pöchhacker & M. Shlesinger (Eds.), *The Interpreting Studies Reader* (pp. 300-311). London and New York: Routledge.

Schmid, H. 1994. *Probabilistic part-of-speech tagging using decision trees*. 10 Mar 2005. Available online: <http://www.ims.uni-stuttgart.de/~schmid/>.

Setton, R. 2001. A methodology for the analysis of interpretation corpora. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century. Challenges and Opportunities. Selected Papers from the 1st Forlì Conference on Interpreting Studies, 9-11 November 2000*. Amsterdam/Philadelphia: John Benjamins.

——. (forthcoming). Corpus-based interpretation studies (CIS): Reflections and prospects. *Corpus-based Translation Studies: Research and Applications*. Paper delivered at *Symposium on Corpus-based Translation Studies: Research and Applications*, Pretoria, 22-25 Jul 2003.

Shlesinger, M. 1989. Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral literate continuum. Unpublished M.A. Thesis, Tel Aviv University, Tel Aviv.

——. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta*, 43/4, 486-493.

——. 2009. Towards a definition of interpretese: An intermodal, corpus-based study. In H. Gyde, A. Chesterman & H. Gerzymisch-Arbogast (Eds.), *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile* (pp. 237-253). Amsterdam/Philadelphia: John Benjamins.

Straniero Sergio, F. 2007. *Talkshow Interpreting. La mediazione linguistica nella comunicazione spettacolare*. Trieste: EUT.

Stubbs, M. 1986. Lexical density: A technique and some findings. In M. Coulthard (Ed.), *Talking About Text. Discourse Analysis* (pp. 27-42). Birmingham: University of Birmingham.

——. 1996. *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford and Cambridge, MA: Blackwell.

Thompson, P. 2005. Spoken language corpora. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 59-70). Oxford: Oxbow Books. Available online: <http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm>.

Ure, J. 1971. Lexical density and register differentiation. In G. E. Perren & J. L. M. Trim (Eds.), *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics* (pp. 443-452). Cambridge: Cambridge University Press.

van Rooy, B. 2005. The feasibility of simultaneous interpreting in university classrooms. *Southern African Linguistics and Applied Language Studies*, 23/1, 81-90.

Vuorikoski, A. R. 2004. *A Voice of its Citizens or a Modern Tower of Babel? The Quality of Interpreting as a Function of Political Rhetoric in the European Parliament*. Tampere: Tampere University Press.

Wallmach, K. 2006. Is South Africa a role model for other multilingual countries? A translator's perspective. *The Linguist*, 45/5. Available online: <http://www.multilingua.co.za/pdfs/Wallmach_2006_The_Linguist.pdf>.

Zanettin, F. 1998. Bilingual comparable corpora and the training of translators. *META*, 43/4, 616-630.

ANNALISA SANDRELLI
FACULTY OF INTERPRETING AND TRANSLATION,
LIBERA UNIVERSITÀ SAN PIO V (LUSPIO), ROME, ITALY.
E-MAIL: <ANNALISA.SANDRELLI@LUSPIO.IT>

CLAUDIO BENDAZZOLI
DEPT. OF INTERDISCIPLINARY STUDIES ON TRANSLATION,
LANGUAGES AND CULTURES,
UNIVERSITY OF BOLOGNA AT FORLÌ, ITALY.
E-MAIL: <CBENDAZZOLI@SSLMIT.UNIBO.IT>

&

MARIACHIARA RUSSO
DEPT. OF INTERDISCIPLINARY STUDIES ON TRANSLATION,
LANGUAGES AND CULTURES,
UNIVERSITY OF BOLOGNA AT FORLÌ, ITALY.
E-MAIL: <RUSSO@SSLMIT.UNIBO.IT>

ISSN 0970-9819

IJT International Journal of Translation

(*A Half-Yearly Review of Translation Studies*)

Vol. 22 No. 1-2 Jan-Dec 2010

Special Issue on:
Translation and Corpus Linguistics - Part II

Guest Editors
María Jesús Blasco Mayor
José Manuel Martínez Martínez
Universitat Jaume I, Castellón, Spain

Founder Editor
Ujjal Singh Bahri

Editors
Harpreet Kaur Bahri
Deepinder Singh Bahri

BAHRI PUBLICATIONS (2010)

ISSN 0970-9819

International Journal of Translation

Vol. 22 **No. 1-2** **Jan-Dec 2010**

International Journal of Translation – IJT, (started in 1989) is a peer-reviewed international journal, which is published twice a year, in March and September. It publishes original research papers related to TRANSLATION STUDIES.

The views expressed herein are those of the authors. *International Journal of Translation* reserves the right to edit the material.

Editors:
Harpreet Kaur Bahri
Deepinder Singh Bahri

Subscription (for 2010):
India: Rs. 699
Rest of the world: US$ 120

Payments can be made by Cheques/Bank Drafts/International Money Orders drawn in the name of BAHRI PUBLICATIONS, NEW DELHI and sent to:

BAHRI PUBLICATIONS
1749A/5, 1st Floor, Gobindpuri Extension,
Kalkaji, New Delhi 110019
Telephones: 011-65810766, (0) 9811204673, (0) 9212794543
E-mails: <bahrius@vsnl.com>; <bahripublications@yahoo.com>
Website: <http://www.bahripublications.in>

Printed & Published by Deepinder Singh Bahri, for and on behalf of **BAHRI PUBLICATIONS**, New Delhi.

International Journal of Translation

Vol. 22 No. 1-2 Jan-Dec 2010

CONTENTS

This volume is part of ongoing work by the research group ECPC under the R&D contract "Ampliación y profundización de ECPC y de CONCECPC 1.0: avances teórico-descriptivos e innovaciones tecnológicas" (Ref. FFI2008-01610) funded by the Spanish Ministry of Science and Innovation, coordinated by Prof María Calzada Pérez.

Editorial

Special Issue on Translation and Corpus Linguistics (Part II)

The special editors of this volume would like to thank *International Journal of Translation* – IJT for inviting us to edit the present issue, a sequel of the first issue edited by María Calzada Pérez and Noemí Marín Cucala. We are also greatly indebted to Carmen Pérez Basanta and María Calzada Pérez, Chairs of the International Seminar whose papers are published here, for kindly giving us the opportunity to present this monograph of selected papers.

The papers included in this volume were presented at the International Seminar *"New Trends in Corpus Linguistics for Language Teaching and Translation Studies. In Honour of John Sinclair,"*[1] held on 23-25 September 2008 at the University of Granada, Spain. Considered by many as one of the fathers of Corpus Linguistics, John Sinclair, looking back on the evolution of the discipline, wrote:

> Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered possible but still lunatic. Today it is very popular. (Sinclair 1991:1)

Almost twenty years later, it is not only popular but most disciplines related to language have steeped themselves in Corpus Linguistics thanks to the work developed by this "giant of English Language Studies."[2] Translation Studies are among them. The clearest proof is the selection of papers for this issue. To paraphrase Newton, today more than ever we can say that if we have seen a little further it is by standing on the shoulders of giants such as John Sinclair.

The compilation presented here offers a broad range of topics stemming from Corpus Linguistics (CL) and its multifarious applications by the Translation Studies (TS) research community. It opens with two different approaches to corpus methodology: the first paper, 'Towards a Science of Corpus Annotation: A New

AUTOCERTIFICAZIONE

I sottoscritti ANNALISA SANDRELLI, CLAUDIO BENDAZZOLI e MARIACHIARA RUSSO, autori del seguente articolo:

"European Parliament Interpreting Corpus (EPIC): Methodological issues and preliminary results on lexical patterns in simultaneous interpreting", *International Journal of Translation*, vol 22, No. 1-2, Jan-Dec 2010: 165-203.

DICHIARANO

di aver elaborato congiuntamente l'articolo di cui sopra, ma di essere singolarmente gli autori delle seguenti sezioni:
- Annalisa Sandrelli §2.3, §3.1 e §3.3;
- Claudio Bendazzoli §1, §2.1, §2.2 e §3.4;
- Mariachiara Russo l'Introduzione e §3.2.

Tutti e tre sono autori delle Conclusioni.

Per un errore materiale da parte degli editori, non è stata riprodotta la seguente nota all'inizio dell'articolo:

> Although the present article is the result of a joint effort, Annalisa Sandrelli wrote §2.3 §3.1 and §3.3; Claudio Bendazzoli wrote §1, §2.1, §2.2 and §3.4; Mariachiara Russo wrote the Introduction and §3.2. The Conclusions were jointly drafted.

Luogo e data

Forlì 18 ottobre 2010

I dichiaranti

Annalisa Sandrelli

Claudio Bendazzoli

Mariachiara Russo

DICHIARAZIONI SOSTITUTIVE DI CERTIFICAZIONI
DICHIARAZIONI SOSTITUTIVE DI ATTI DI NOTORIETA'
(ART. 46 e 47 D.P.R. 445 DEL 28/12/2000)

I sottoscritti, consapevoli delle sanzioni penali, nel caso di dichiarazioni non veritiere, di formazione o uso di atti falsi, richiamate dall'art. 76 del D.P.R. 445 del 28 dicembre 2000, dichiarano che quanto sopra riportato corrisponde al vero.

Luogo e data

Forlì 18 ottobre 2010

I dichiaranti

Annalisa Sandrelli

Claudio Bendazzoli

Mariachiara Russo