Lexical patterns in simultaneous interpreting
a preliminary investigation of EPIC (European Parliament Interpreting Corpus)

Sandrelli Annalisa & Bendazzoli Claudio[1]
Directionality Research Group[2]
Department of Interdisciplinary Studies in Translation, Languages and Cultures (SITLeC)
University of Bologna at Forlì
Annalisa.Sandrelli@unibo.it cbendazzoli@yahoo.com

1 Introduction

The present paper presents the first results of a preliminary investigation of EPIC, the European Parliament Interpreting Corpus, that is being compiled in the Department of Interdisciplinary Studies in Translation, Languages and Cultures (SITLeC) of the University of Bologna at Forlì. EPIC is an open, parallel, trilingual (Italian, English and Spanish) corpus of European Parliament speeches and their corresponding simultaneous interpretations (Monti et al. forthcoming, Bendazzoli & Sandrelli forthcoming, Bendazzoli et al. 2004). The main reason for the creation of the corpus was to collect a large sample of homogeneous interpreting data in order to overcome the main obstacle hampering research on simultaneous interpreting, that is, the lack of access to reliable data in sufficient quantities (see Cencini 2002, Monti et al. forthcoming, Bendazzoli & Sandrelli forthcoming for a discussion of the methodological and practical problems of interpreting research). As is described in more detail in §2, several European Parliament sittings were recorded along with the performances of the interpreters working in the English, Italian and Spanish booths. The highly formal and institutionalised setting of the European Parliament ensures the homogeneity of the source speeches, whereas the strict interpreter selection process guarantees similar levels of expertise in all of the interpreters working there, and consequently a high degree of homogeneity in the target (interpreted) speeches as well.

EPIC was created with a view to studying the effects of directionality in simultaneous interpreting, i.e. whether interpreters use different strategies when interpreting between cognate languages and between languages belonging to different language families. The present study is a first attempt to explore part of the data collected until now, starting with an overview of general lexical patterns in the corpus.

---

[1] Although the present paper is the product of a joint effort, Annalisa Sandrelli can be identified as the author of §1 and 3, while Claudio Bendazzoli is the author of §2 and 4. The conclusions (§5) have been jointly drafted.

[2] The other members of the Directionality Research Group are Mariachiara Russo, Cristina Monti, Marco Baroni, Elio Ballardini, Silvia Bernardini, Gabriele Mack and Peter Mead. The EPIC web designers are Lorenzo Piccioni and Eros Zanchetta.

Our starting point is Laviosa's work on lexical density in the Translational English Corpus, or TEC (Laviosa 1998), which comprises both translated narrative prose (into English from a number of European languages) and original narrative texts written in English. Laviosa (1998: 563) found that translated texts in TEC display four main lexical patterns:

"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);

iv) The list head of translated texts contains fewer lemmas."

We aim to investigate whether the first three of the above patterns apply only to (written) translated texts or whether similar patterns can be found in our corpus of (spoken) interpreted speeches as well. The fourth finding on the number of lemmas in the list head of translated texts was excluded from the aims of the present study because tagging and lemmatisation of our corpus are still imperfect at this stage and lemmatised lists would not have been entirely reliable.[3]

Furthermore, since all the EPIC source language speeches in English, Italian and Spanish have been interpreted into the other two languages, we aim to verify whether there are differences in lexical density according to language pair and language direction: we hypothesise that there will be differences depending on the language combination (two Romance languages or one Romance language and a Germanic language). However, it must be pointed out that the materials under study in this article include only the English and Italian source and target speeches. The Spanish source speeches and the speeches interpreted into Spanish will be studied in a future stage of the project.

Section 2 gives a detailed description of the materials under analysis. Section 3 illustrates the methodology followed to verify Laviosa's results on lexical density and presents our findings. Section 4 examines the list heads of EPIC source and target speeches and section 5 presents our conclusions and directions for future research on this issue.

2 Corpus description

As was mentioned in §1, the material analysed in the present study is a part of the European Parliament Interpreting Corpus (EPIC), which is described in the present section.

[3] Indeed, the creation of several training sub-corpora is the next step in the development of EPIC, to improve the reliability of the tagging.

In 2004 several European Parliament plenary sittings were recorded off the news channel EbS (Europe by Satellite), using four TV sets and video-recorders with satellite decoders. By selecting different audio channels, it was possible to record the original speakers and the interpreters working in the various booths (in our case, Italian, English and Spanish). All the material thus obtained is being digitised and edited by using dedicated software in order to create a multimedia archive (described in detail in Bendazzoli & Sandrelli forthcoming). The EPIC archive includes digital video clips of the source speeches in English, Italian and Spanish and the audio clips of the two corresponding interpreted versions. The material currently available in digital form comprises a large part of the EP debates held in February and July 2004, totalling about 600 video and audio clips. Digitisation and editing are continuing to further expand the archive.

The clips thus obtained are transcribed, POS-tagged and lemmatised to create the EPIC corpus. This is done by using existing taggers, that is *Treetagger* (Schmid 1994) for English, *Freeling* (Carreras et al. 2004) for Spanish and the combination of taggers suggested by Baroni et al. (2004) for Italian. At the time of writing, 357 clips have been transcribed and tagged, corresponding to about 21 hours of spoken material.

Currently, the EPIC corpus is made up of nine sub-corpora, which can be queried individually. There are three sub-corpora of source speeches in the three languages under study (named org-en, org-it, and org-es) and 6 sub-corpora of (interpreted) target speeches (indicated as "int" followed by the language direction, e.g. int-en-it for English into Italian). Thus, all the combinations and directions of the three languages are covered.

The transcripts feature a header containing linguistic and extra-linguistic information about the speech and the speaker. The information recorded in the various header fields has been used to set the search filters available in a dedicated EPIC web interface.[4] The latter also provides information about transcription criteria and conventions, the EPIC multimedia archive and some general information about EP debates, including the rules for the allocation of speaking time.

All the tagged material has been encoded by using 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. This makes it possible to formulate simple and advanced queries in the CQP language of *CWB* through the web interface, and to restrict queries on the basis of the search filters, i.e. the structural attributes. An example of the tagged and encoded corpus can be seen in figure 1 below, in which the XML attributes are followed by a first column which contains the tokens, a second column with the tags, a third column of lemmas, and a

[4] The EPIC web interface is available at http://sslmitdev-online.sslmit.unibo.it/corpora/corpora.php). The website also hosts other corpora projects, as well as a number of useful resources for linguists and terminologists.

fourth and final column with a transcript of how the words were actually uttered, including any disfluencies (e.g. *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>

Figure 1 Example of an EPIC transcript

The interface features several speaker-related and speech-related search filters. Examples of the former include gender, country, political function, and so on, whereas examples of the latter are duration, speech length, pace of delivery, mode of delivery, etc. In particular, duration and speech length were classified as short, medium or long, and speed of delivery (calculated as the number of words per minute) as low, medium or high, according to the following values:

duration	short < 2 minutes medium 2-6 minutes long > 6 minutes
text length	short < 300 words medium 301-1000 words long > 1000 words
speed of delivery	low < 130 words per minute (w/m) medium 131-160 w/m high > 160

Table 1 Values assigned to duration, text length and speed in EPIC transcripts

It is worth specifying that the above reference values assigned to each label were established on the basis of the current material in the corpus. In other words, they can only be considered valid within the specific context of EP debates, in which 150 w/m, for instance, can be considered an "ordinary" speed of delivery (see Monti et al. forthcoming and Bendazzoli & Sandrelli forthcoming).

As regards the mode of delivery, when the speakers did not glance at any notes, the speeches were classified as *impromptu*, whereas when they were clearly seen to be reading a script, the speech was classified as *read*. The *mixed* label describes situations in which speakers kept switching between not using notes and reading fragments of a prepared script. Clearly, this is a simplified classification used to categorise the countless varieties along the written-to-spoken continuum (Nencioni 1976). This information may prove useful in future studies of interpreters' strategies, since the mode of delivery is a significant variable affecting comprehension. Déjean Le Féal (1982) explains that

impromptu speeches are easier to understand (both for the audience and the interpreter), because of a number of features pertaining to sentence segmentation, prosody and degree of redundancy.[5] Table 2 presents an outline of the current size and composition of EPIC. The sub-corpora in bold are the ones included in the present study.

sub-corpus	n. of speeches	total word count	% of EPIC
Org-en	81	42705	24
Org-it	17	6765	3.8
Org-es	21	14468	8.2
Int-it-en	17	6708	3.8
Int-es-en	21	12995	7.3
Int-en-it	81	35765	20.1
Int-es-it	21	12833	7.2
Int-en-es	81	38435	21.6
Int-it-es	17	7073	4
TOTAL	**357**	**177748**	100

Table 2 Composition of EPIC

The following subsections describe the main features of the 6 sub-corpora in question.

2.1 Source speeches

2.1.1 Description of org-en

The sub-corpus named org-en, that is the source speeches delivered in English, is the largest one in EPIC, accounting for almost 24% of the overall word count (see table 2 above). It comprises 81 speeches, 3 of which delivered by non-native speakers (from Denmark, the Netherlands and Portugal, respectively). 35 speeches were 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 of the speeches are delivered by Members of the European Parliament (42, as well as 13 speeches by the EP President and 1 by a Vice-President), but there are also some speeches made by European Commissioners (18) and Ministers of the European Council (7). As regards the speeches delivered by MEPs, speech distribution by speakers' political group can be seen in figure 2 below.

[5] In terms of sentence segmentation, impromptu speeches are usually made up of shorter fragments than read speeches, because speakers plan their sentences as they go along. These shorter chunks are easier to process for listeners. Moreover, intonation patterns, which guide listeners to meaning, are more marked in impromptu speeches; besides, the pace of delivery is often lower than in read speeches. Finally, redundancy is generally higher than in read speeches, both linguistically and content-wise, because the text is less tightly-structured and because speakers tend to tailor it to the audience's perceived degree of receptivity (Déjean Le Féal 1982).

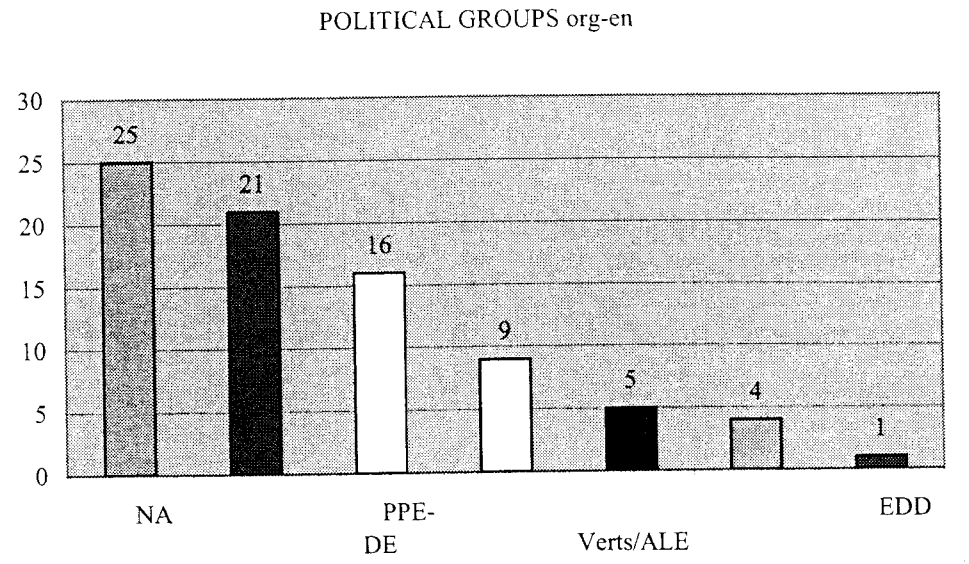

Figure 2 MEP speakers by political groups in org-en

Turning to the characteristics of the English source speeches, more than half were read from a written script (43 out of 81), whereas just over one fourth (24) were delivered impromptu. The remaining speeches (14) were delivered in a mixture of read and impromptu mode.

In terms of duration, half of the speeches are medium (40), that is they last between 2 and 6 minutes. 28 speeches are short, and only 13 were classified as long. The average duration is thus around 3 min 30 secs. Clearly, text length (i.e. word count) reflects similar patterns, in that over half (44) of the English source speeches are of medium length, 27 speeches are short and only 10 speeches are long.[6]

Interestingly, looking at speed, the speeches delivered at a fast pace (34) are almost as many as those given 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 range from politics to health to economics, with political speeches taking the lion's share, as can be seen in figure 3:

[6] The slight discrepancy between the figures related to duration and text length is due to the fact that the number of words in a speech depends not just on its duration but also on the speaker's delivery rate.

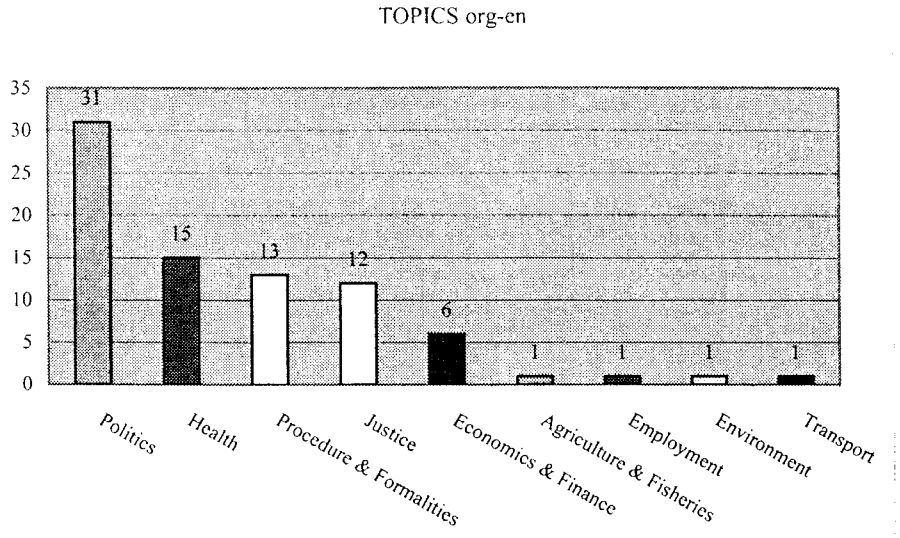

Figure 3 Topics discussed in org-en speeches

2.1.2 Description of org-it

This EPIC sub-corpus comprises 17 Italian source speeches delivered by native Italian speakers. These are all MEPs, 14 men and 3 women, belonging to different political groups, as shown in figure 4 below:

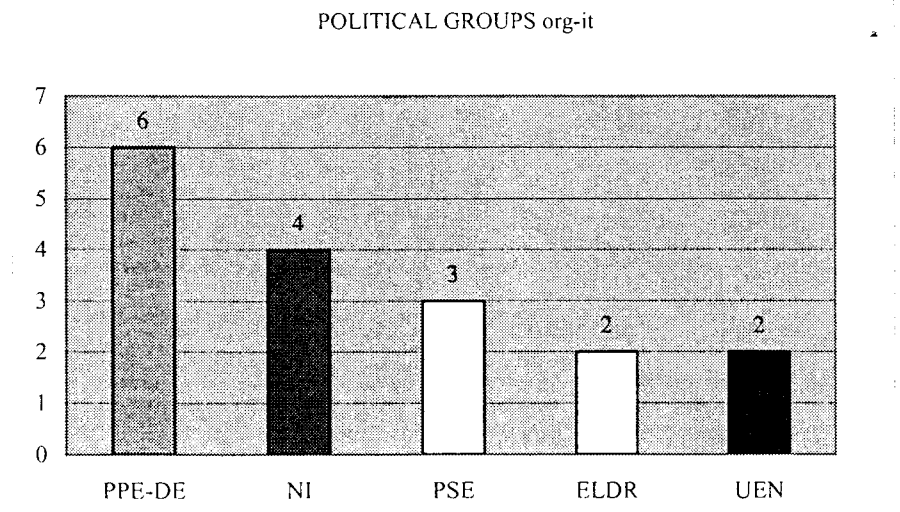

Figure 4 MEP speakers by political groups in org-it

8 speeches were read out of a written text, 6 were delivered off-the-cuff, while 3 were delivered in a mixed mode. In terms of duration, 13 speeches were classified as medium, while only 4 as short. The overall duration of Italian source speeches amounts to almost 50 minutes, with an average duration of 3 minutes per speech.

This sub-corpus comprises 6765 words in total (see table 2 in §2.). There are 10 medium-length and 7 short speeches, with an average count of about 400 words per speech. Speed of delivery is low in 11 speeches and medium in 6 speeches. On average, this set of Italian speeches was delivered at a speed of about 130 words per minute.

Topics vary considerably in EP debates. The Italian source speeches are no exception, as shown in figure 5:

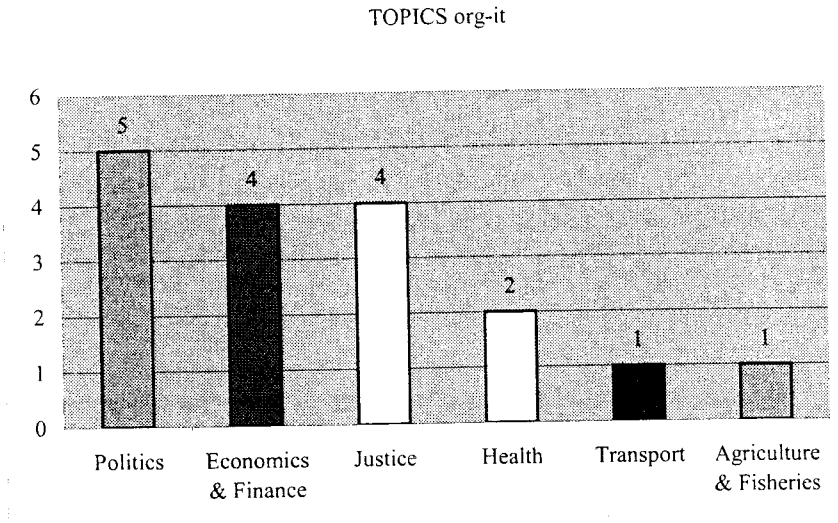

Figure 5 Topics discussed in org-it speeches

2.2 Sub-corpora of target (interpreted) speeches

2.2.1 Speeches interpreted into English

The two sub-corpora of speeches interpreted into English are int-it-en and int-es-en (from Italian and Spanish, respectively).

The sub-corpus of English target speeches interpreted from Italian source speeches is the smallest one in EPIC, together with, obviously, the collection of its Italian source speeches (org-it; see §2.1.2). It comprises 17 target speeches delivered by 8 male interpreters and 9 female interpreters, 16 of them native speakers and one non-native speaker. The average speech length is 387.5 words, that is, slightly shorter than the corresponding source language speeches. As regards speed, 8 speeches were delivered at low speed, 8 at medium speed and 1 at high speed. The average is 132.2 w/m, that is, slightly faster than the average for the source language speeches (again see §2.1.2, above).

The sub-corpus of speeches interpreted from Spanish into English is made up of 21 speeches. As has already been pointed out, the Spanish source texts are not included in the present study; this subsection briefly presents the main features of this group of speeches which were then interpreted

into English and Italian. In terms of topic, once again the largest group is that of political speeches (10), followed by speeches on justice (5) and economics and finance (3).

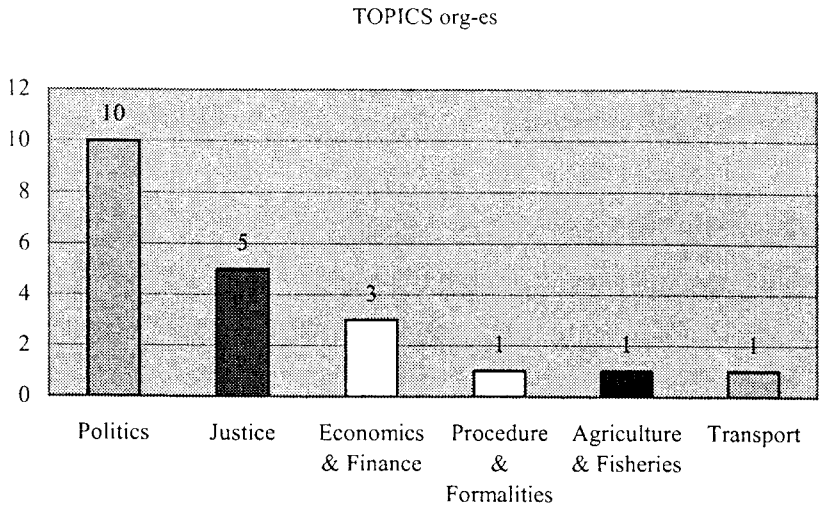

Figure 6 Topics discussed in org-es speeches

The majority of speeches (13) are in the medium duration category, with 3 speeches classified as long and the remaining 5 as short. The average duration of the Spanish source speeches is about 4 minutes 40 secs. 5 speeches were delivered impromptu, 7 in a mixed mode and 9 were read. Turning to the English interpreters who had to translate this particular subset of speeches, there were 16 men and 5 women, all of them native speakers. Their pace of delivery was, on average, 136.2 w/m. More specifically, 4 speeches were delivered at high speed, 9 at low speed and 8 at medium speed. In terms of text length, the interpreted versions are mostly medium (13), with only 5 short speeches and 3 long ones: the average length in the int-es-en sub-corpus is 608.4 words.

2.2.2 Speeches interpreted into Italian

EPIC comprises two sub-corpora of speeches interpreted into Italian, namely int-en-it (i.e. interpretations from English into Italian) and int-es-it (i.e. interpretations from Spanish into Italian). The int-en-it sub-corpus is the largest one among the collections of target speeches, since the source texts come from the large org-en sub-corpus (see 2.1.1). The vast majority of interpreters were women (68 vs. 13 men). The average speed of delivery is 123.7 w/m per minute (lower than that of the English source speeches), and the average length of each interpreted speech is 428.5 words. On the other hand, the int-es-it sub-corpus is made up of 21 speeches interpreted from Spanish into Italian (see 2.2.1 on the main characteristics of the Spanish source speeches), for a total of 12830 words. Interpreters working in this direction are all women. Their average pace of delivery was 124.5 words per minute and the average speech length is about 594 words.

3 Lexical density

After describing extensively the 6 sub-corpora under study, let us go back to our original aims as they were stated in the Introduction (§1). The first objective is to investigate lexical density in order to verify whether it is lower in the sub-corpora of interpreted speeches than in the sub-corpora of source speeches, in other words to confirm Laviosa's findings on translated texts in TEC. Laviosa (1998: 565) defines lexical density as follows:

"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."

Before lexical density can be calculated for each of our sub-corpora, an operational definition of function words and lexical words is needed. Reference is here made to the distinction between **closed-class** and **open-class** parts of speech made by Jurafsky & Martin (2004: 3): "Closed classes are those that have relatively fixed membership. For example, prepositions are a closed class because there is a fixed set of them in English; new prepositions are rarely coined. By contrast nouns and verbs are open classes because new nouns and verbs are continually coined or borrowed from other languages [...]". Closed-class words are **function words**, whereas open-class words are **lexical words**. The main types of function words are prepositions, determiners, pronouns, conjunctions, particles, numerals, interjections, negatives, greetings, and politeness markers. The main groups of lexical words are nouns, adjectives, verbs and adverbs. We used this categorisation to compile the lists of function words and lexical words in the sub-corpora of English and Italian source and target (interpreted) speeches. However, it must be stressed that this taxonomy is not as impermeable as it may appear: "Although they have deceptively specific labels, the word classes tend in fact to be rather heterogeneous, if not problematic categories. There is nothing sacrosanct about the traditional parts-of-speech-classification [...]" (Quirk et al. 1985: 73). Indeed, both in English and Italian, even prepositions may be divided into a 'closed' set and a more 'open' set of prepositional phrases (preposition + noun + preposition), which are the more creative subgroup (Quirk et al. 1985: 72; Dardano and Trifone 1989: 396).

In particular, verbs may be classified as primary, modal and full verbs (the first two belonging to the function word category and the third one to the lexical word category), or, as we chose to do, simply as verbs (lexical words).

The class of adverbs is particularly varied, including adverbs with an adjectival base (with an *–ly* suffix in English and a *–mente* suffix in Italian) and others belonging to a closed class, such as *here – qui*, *now – ora*, etc. Moreover, some adverbs may be classified as conjunctions or as prepositions

according to the position and function they play in a given sentence. For example, the Italian word "*perché*" may be used as an interrogative adverb or as a conjunction; the same applies to the temporal adverb *quando* which is also used as a conjunction. Similarly, the English adverbs *around* and *behind* may also function as prepositions, whereas the word *before* may be used as an adverb, a preposition or a conjunction.

Other problematic categories are those words which may function either as adjectives (lexical words) or pronouns (function words). Examples include demonstratives, possessives, distributives, quantifiers, to name but a few.

Since the tagging of EPIC is still slightly inaccurate (see §1), all the problematic cases were analysed in KWIC view (available through the web interface), and the resulting occurrences were manually counted and assigned to the relevant lists of lexical or function words.

In addition, two specific characteristics of EPIC had to be taken into account before lexical density could be calculated. Firstly, dates and figures are fully spelt out in our transcripts to prevent problems in the tagging process. In practical terms this means, for example, that the same figure accounts for 3 tokens in English and only 1 in Italian: *two hundred thousand* vs. *duecentomila*. However, this structural difference only seems to affect the overall word count very marginally. As can be seen in table 3, cardinal numbers accounted for 1.87% of the total word count for the English source speeches; in the Italian interpreted versions of the same speeches (int-en-it) the percentage goes down to 1.28%, which may be attributed partly to the structural difference referred to above, and partly to a few omissions by the interpreters. Going in the opposite direction, that is, from Italian into English, the percentage slightly increases, from 0.88% to 1.05%. These figures (-0.59% in the English into Italian direction and +0.17% in the Italian into English direction) seem to indicate that this particular structural difference may be disregarded when calculating lexical density.

	cardinal numbers	**% of sub-corpus**
org-en	800	1.87%
int-it-en	71	1.05%
int-es-en	267	2.05%
org-it	60	0.88%
int-en-it	458	1.28%
int-es-it	206	1.6%

Table 3 Incidence of cardinal numbers in the 6 wordlists

Another aspect that was taken into account is truncated words, a characteristic feature of EPIC speeches. During the transcription process, words which were not fully uttered by speakers and

interpreters were orthographically transcribed, adding a dash at the end of each word (for example *thes-*). Truncated words were all counted as lexical words, because in most cases they were immediately followed by the full word, which was nearly always a lexical word. As can be seen in table 4 below, the percentage of truncated words is fairly low in all the sub-corpora under study and can therefore be disregarded in the calculation of lexical density. However, it is interesting to note that in the Italian source speeches the incidence of truncated words is noticeably lower than in the sub-corpus of English source speeches, which seems to reflect more laborious speaking patterns on the part of English native speakers.[7] On the other hand, the incidence of truncated words is very similar in all interpreted speeches, with the exception of the English into Italian direction, where it is lower. This seems to indicate better control of their own target language production by this group of interpreters.

sub-corpus	number of truncated words	%
org-en	391	0.9
int-it-en	68	1.0
int-es-en	120	0.9
org-it	29	0.4
int-en-it	219	0.6
int-es-it	116	0.9

Table 4 Incidence of truncated words

Bearing in mind all of the above provisos, the steps taken to create the lists of function words and lexical words were as follows. As was briefly explained in §2, EPIC has been encoded by using the *IMS Corpus Work Bench – CWB* (Christ 1994). Therefore, a relevant command was issued in the command line of a machine connected to a dedicated Unix server to extract all the tokens and corresponding tags from the 6 sub-corpora. The function words were selected from the 6 files thus created on the basis of their tags. The 6 lists of function words thus obtained were then manually "cleaned" and any mistakes were corrected. This was a time-consuming but necessary step, which enabled us to calculate the overall number of function words and lexical words for each sub-corpus, as can be seen in table 5:

sub-corpus	total running words	lexical words	function words	lexical density
org-en	42705	24475	18230	57.311790188
int-it-en	6708	3872	2836	57.722122838
int-es-en	12995	7419	5576	57.0911889188

[7] A possible research question for a future study could be whether the higher percentage of truncated words in the English source speeches is also accompanied by a higher percentage of hesitations (indicated by empty and filled pauses in the transcripts), thus confirming a marked difference in speech planning patterns between the two groups of speakers.

org-it	6765	3997	2768	59.08351811
int-en-it	35765	21209	14556	59.30099259
int-es-it	12833	7452	5381	58.06904075

Table 5 Lexical density in the sub-corpora under analysis

Finally, lexical density was calculated for each sub-corpus. The percentages thus obtained are commented in §3.1 and 3.2 below.

3.1 Lexical density in the English sub-corpora

The effect noted by Laviosa (1998) in translated texts, i.e. a highly significantly lower lexical density than in original texts written in English, is not confirmed.[8] In fact, there is little variation in lexical density when interpreted speeches are compared with original English speeches. 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 higher than in the original English speeches (+0.41033265).

3.2 Lexical density in the Italian sub-corpora

The effect on lexical density noted by Laviosa is not confirmed in the group of speeches interpreted from English into Italian, in which lexical density is slightly higher than in the original Italian speeches (+0.21747448). By contrast, in the subset of speeches interpreted from Spanish into Italian lexical density decreases more substantially (-1.01447736), but it is difficult to say at this stage whether this difference is significant.

The results obtained by analysing lexical density in the 6 sub-corpora are not easy to interpret. The general trend seems to indicate that there is very slight variation in lexical density in simultaneously interpreted texs in comparison with speeches originally produced in English and Italian. This may be due to the specific text production conditions, i.e. the pace of the incoming speech is imposed by the source speaker and the interpreter has to assemble the target speech practically "on-line", chunk by chunk, by selecting and re-arranging information to suit the norms of the target language. The parallel co-existence of source and target speeches and the time constraints under which interpreting is performed may explain why the patterns observed by Laviosa in relation to written texts do not apply. However, it must be noted that the only exception to our findings is the group of speeches interpreted from Spanish into Italian, i.e. the only combination of two Romance languages analysed in the present paper. The importance of this may become clearer after an examination of the list heads, described in §4.

4 List heads

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.[9]

The first 100 occurrences in our frequency lists were selected to create our 6 list heads. The overall word count was calculated for each list head, as well as the percentage of sub-corpus accounted for by high frequency words. The data are presented in tables 6 and 7 below, for English and Italian respectively.

4.1 English 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	22745	53.26	6142	27.0	16603	73.0
int-it-en	3832	57.09	1250	32.6	2582	67.4
int-es-en	7176	55.22	2112	29.4	5064	70.6

Table 6 List head word counts and percentages in original and interpreted English

The data in table 6 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 seem to 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.

As for the distribution of lexical and function words in the list heads, the English source speeches show a lower percentage of lexical words than both interpreted English sub-corpora. This may indicate the interpreters' tendency 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 than speeches originally produced in English. In order to test this hypothesis, it will be necessary to align the interpreted speeches with their source speeches, in other words, to create parallel sub-corpora, as advocated by Shlesinger (1998).[10]

---

[8] See Laviosa (1998: 565) for more details on the statistical significance test used.

[9] This was demonstrated by Laviosa by creating the list heads of the translational and the non-translational components of her corpus (i.e. by selecting the 108 most frequent words in the two frequency lists), and by counting the corresponding occurrences. Then, she calculated the percentage of the corpus (translational and non-translational) represented by the two list heads.

[10] The alignment of our sub-corpora is indeed one our objectives for the future development of EPIC.

4.2 Italian list heads

The results in table 7 regarding Italian source and target speeches do not seem to be in line with Laviosa's findings for written translation:

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-it	3365	49.74	892	26.5	2473	73.5
int-en-it	17353	48.51	4771	27.5	12582	72.5
int-es-it	6264	48.82	1572	25.1	4692	74.9

Table 7 List head word counts and percentages in original and interpreted Italian

The list heads of the speeches interpreted into Italian (from English and Spanish) account for a smaller portion of their respective sub-corpora. In this case, it seems that the nuclei of most frequently used words are more varied in speeches interpreted into Italian than in the source speeches originally delivered in Italian. This is an unexpected finding, which may be related to corpus size. Table 7 (above) shows that the lowest percentage of high frequency words (48.51%) can be found in the largest list head (int-en-it), and the percentage seems to increase as size decreases. A similar trend can be observed in table 6 containing the data on the English list heads. In this case, the smallest proportion of high frequency words is found in the largest list head (org-en), whereas the highest percentage corresponds to the smallest list head (int-it-en). Clearly, this observation is not conclusive. Further investigation is required to confirm that there is an effect of corpus size on the percentage of high frequency words in the list heads.

As for the distribution of lexical and function words in these list heads, the trend observed above in the English interpreted texts, i.e. the higher percentage of lexical words in comparison with the source speeches, is confirmed only in the sub-corpus of speeches interpreted from English into Italian. Indeed, in the latter group of speeches, lexical words account for 27.5% of the list head, in comparison with 26.5% in the set of original Italian speeches. By contrast, in the speeches interpreted from Spanish into Italian the opposite trend can be observed.

The findings illustrated in §4.1 and 4.2 seem contradictory. The data on English source and target speeches are in line with Laviosa's suggestions, whereas the opposite is true of the Italian source and target speeches.

5 Discussion and conclusions

This study is a first attempt to explore the European Parliament Interpreting Corpus by using corpus linguistics techniques and semi-automatic analysis. Given its very preliminary nature, this study has its limits. Firstly, the sub-corpora are of different sizes, with the English source speeches and their interpreted versions into Italian and Spanish accounting for 65.7% of the overall word count of EPIC. As was mentioned in §4.2, corpus size may determine an effect on the composition of the list heads; therefore, the size imbalance needs to be gradually corrected by adding more materials to EPIC, which is an open, expanding corpus. This will enable us to carry out further studies on lexical density and high frequency words so as to confirm or disprove our present conclusions.

Another limitation is the accuracy rate of the taggers, which were designed for written texts and therefore do not work perfectly on the many features of spoken language displayed by EPIC speeches, including false starts, reformulations, truncated words, etc. There are plans to manually correct the tagging of part of the corpus (i.e. to create a training corpus), so as to improve the success rate of our taggers. This will enable us to fully exploit existing techniques for automatic extraction and analysis of corpus data.

The data obtained on lexical patterns in the portion of EPIC under study do not fully confirm Laviosa's conclusions on translational English. On the one hand, lexical density does not seem to be affected by the interpreting process, with the exception of the Spanish into Italian direction. On the other hand, it must be stressed that Laviosa's findings concern translational English only, which may mean that they are not automatically applicable to translational (and interpreted) Italian.

As regards the list heads, i.e. high frequency words, the English interpreted speeches display less lexical variety than the original English speeches, thus confirming Laviosa's observations. However, the opposite is true in the Italian interpreted speeches.

A further observation can be made by looking at the relative weight of lexical and function words in our list heads. This detailed analysis has once again highlighted a difference in the sub-corpus of speeches interpreted from Spanish into Italian. The reasons for the different "behaviour" of this particular subset of speeches are unclear and will require further investigation. However, it is worth observing that this is the only group involving a combination of two Romance languages. This seems to suggest that language pair may play a significant role in simultaneous interpreting, as has often been claimed (among others, Snelling 1992, Viezzi 1999, Falbo et al. 1999, Kelly et al. 2003, and Donovan 2004). This tentative conclusion may be tested when the sub-corpora of Spanish source and target speeches (org-es, int-en-es, int-it-es) are processed for the same type of analysis. Furthermore, when EPIC is fully aligned, it will be possible to draw comparisons between source

speeches and their interpretations, that is to explore it not only as a comparable corpus but as a parallel corpus as well.

References

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. and M. Mazzoleni (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, and R. Silva (eds.) with the collaboration of C. Pereira, F. Carvalho, M. Lopes, M. Catarino and S. Barros *Proceedings of the 4th International Conference on Language Resources and Evaluation*, ELRA 5, 1771-1774.

Bendazzoli C. & A. Sandrelli (forthcoming) An approach to corpus-based interpreting studies: developing EPIC (European Parliament Interpreting Corpus), in *Proceedings of MuTra. Challenges of Multidimensional Translation.* EU High-Level Conference Series, Saarbruecken 2-6 May 2005, (Manchester: St Jerome Publishing).

Bendazzoli, C., Monti, C., Sandrelli, A., Russo, M., Baroni, M., Bernardini, S., Mack, G., Ballardini, E. and P. Mead (2004) Towards the creation of an electronic corpus to study directionality in simultaneous interpreting, in N. Oostdijk, G. Kristoffersen, and G. Sampson (eds.) *Compiling and Processing Spoken Language Corpora*, LREC 2004 Satellite Workshop, Fourth International Conference on Language Resources and Evaluation, 24 May 2004, 33-39.

Carreras, X., Chao I., Padró, L. and M., Padró (2004) Freeling: an open-source suite of language analyzers, in M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva (eds.), with the collaboration of C. Pereira, F. Carvalho, M. Lopes, M. Catarino and S. Barros *Proceedings of the 4th International Conference on Language Resources and Evaluation*, ELRA 1, 239-242.

Cencini, M. (2002) On the Importance of an Encoding Standard for Corpus-based Interpreting Studies. Extending the TEI Scheme. *inTRAlinea*, Special Issue CULT2K. [available from: http://www.intralinea.it/specials/eng_open1.php?id=P107/]

Christ, O. (1994) A Modular and Flexible Architecture for an Integrated Corpus Query System. *COMPLEX '94*, Budapest 1994.

Dardano, M. and Trifone, P. (1989) *Grammatica italiana con nozioni di linguistica* (Bologna: Zanichelli Editore).

Déjean Le Féal, K. (1982) Why impromptu speech is easy to understand, in N. E. Enkvist (ed.) *Impromptu Speech. A Symposium* (Åbo: Åbo Akademi), 221-239.

Donovan, C. (2004) European Masters Project Group: Teaching simultaneous interpretation into a B language: Preliminary findings. *Interpreting*, 6-2, 205 -216.

Falbo C., Straniero Sergio F. and M., Russo. (1999) (a cura di) *Interpretazione simultanea e consecutiva. Problemi teorici e metodologie didattiche* (Milano: Hoepli).

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

Jurafsky, D. and Martin, J. H. (2004) Word classes and part-of-speech tagging, revised 2004 version, original chapter in *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Upper Saddle River, Prentice Hall), 2000. [available from http://www.cs.colorado.edu/~martin/slp.html]

Kelly, D., Martin, A., Nobs M., Sánchez, D. and C. Way (2003) *La direccionalidad en traducción e interpretación: perspectivas teóricas, profesionales y didácticas* (Granada: Editorial Atrio).

Laviosa, S. (1998) Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* XLIII: 4, 557-570.

Monti, C., Bendazzoli, C., Sandrelli, A. and M., Russo (forthcoming) Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus). *Meta*.

Nencioni, G. (1976) Parlato-parlato, parlato-scritto, parlato-recitato. *Strumenti critici*, XXIX, febbraio (Torino: Einaudi).

Quirk, R., Greenbaum, S., Leech, G. and J., Svartvik (1985) *A comprehensive grammar of the English language* (London/New York: Longman), indexed by David Crystal, 15th impression.

Shlesinger, M. (1998) Corpus-based interpreting studies. *Meta* XLIII: 4, 486-493.

Snelling D. (1992) *Strategies for Simultaneous Interpreting. From Romance Languages into English*, Università degli Studi di Trieste, Scuola Superiore di lingue Moderne per Interpreti e Traduttori (Udine: Campanotto Editore).

Viezzi, M. (1999) Interpretazione simultanea: attività specifica per coppie di lingue?. *Settentrione* 11: 1, 133-159.

UNIVERSITY OF BIRMINGHAM

Proceedings from Corpus Linguistics 2005

This is a collection of the papers presented at the Corpus Linguistics 2005 conference which was held in Birmingham July 14-17 2005. Some of the papers are either as Word documents or as PDF files.

The proceedings have been divided into 11 subcategories:

Compiling a Corpus

Contrastive Corpus Linguistics

Discourse

Evaluation and Stance

Grammar

Language Learning & Error Analysis through Corpora

Language Processing & Corpus Tools

The Lexicon

Phraseology & Patterns in Language

The Web as a corpus

Spoken Discourse

Compiling a corpus

Rachel Aires, Diana Santos & Sandra Aluisio: "Yes, user!": compiling a corpus according to what the user wants
See "Yes, user !".doc

Latifa Al-Sulaiti and Eric Atwell: Extending the Corpus of Contemporary Arabic
See Extending the Corpus of Contemporary Arabic.doc

Wendy Anderson & Dave Beavan: Internet delivery of time-synchronised multimedia: the SCOTS Projects
See Traditional transcriptions.doc

Caroline Barri re & Akakpo Agbago: Corpus Construction for Terminology
See Terminology.doc

Sara Piccioni: The Lorca corpus at the crossroads of philology and corpus linguistics
See The Lorca corpus at the crossroads of philology and corpus linguistics.doc

Gong Wengao: English in computer-mediated environments: a neglected dimension in large English corpus compilation
See English in computer-mediated environments.pdf

Hilary Nesi, Sheena Gardner, Richard Forsyth, Dawn Hindle, Paul Wickens, Signe Ebeling, Maria Leedham, Paul Thompson and Alois Heuboeck: Towards the compilation of a corpus of assessed student writing
See Towards the compilation of a corpus.doc

Contrastive Corpus Linguistics

Gisle Andersen: Assessing algorithms for automatic extraction of anglicisms in Norwegian texts
See Assessing algorithms.doc

J zsef Andor: A Lexical Semantic-Pragmatic Analysis of the Meaning Potentials of Amplifying Prefixes in English and Hungarian A Corpus-based Case Study of Near Synonymy
See A Corpus-based Case Study of Near Synonymy.doc

Sandrelli Annalisa & Bendazzoli Claudio: Lexical patterns in simultaneous interpreting: a preliminary investigation of EPIC (European Parliament Interpreting Corpus)
See Lexical patterns in simultaneous interpreting a preliminary investigation of EPIC.doc

Marianna Apidianaki: Translation prediction using word co-occurrence graphs
See Translation prediction using word co-occurrence graphs

Tatjana Bala ic Bulc: Connectors in students' academic writing in two closely related languages
See Connectors in students' academic writing in two closely related languages.doc

Silvia Bernardini & Marco Baroni: Spotting translationese: A corpus-driven approach using support vector machines
See Spotting translationese.doc

Gabriela Castelo Branco Ribeiro & Maria Carmelita Padua Dias: Two corpus-based studies about the translation of adjectives in English and Brazilian Portuguese

Wallace Chen: Patterns of Connectors in the English-Chinese Parallel Corpus of Popular Science Texts

Debbie Elliott: Using corpora to automatically detect untranslated and outrageous words in machine translation output

Ana Frankenberg-Garcia: A corpus-based study of loan words in original and translated texts
See A corpus-based study of loan words.doc

Randall L. Jones : Analysis of lexical correspondence in an English-German parallel corpus

Zhenglin Jin & Caroline Barriere: Exploring sentence variations in bilingual corpora
See Exploring sentence variations with bilingual corpora.doc

Tony McEnery and Richard Xiao: Passive constructions in English and Chinese: A contrastive and translation study
See Passive constructions in English and Chinese.doc

Stella Neumann and Silvia Hansen-Schirra : The CroCo Project: Crosslinguistic corpora for the investigation of explicitation in translations
See The CroCo Project.pdf

Pablo Romero Fresco: The translation of phraseology in a parallel (English-Spanish) audiovisual corpus.
See The translation of phraseology in a parallel.doc

Doaa A. Samy: Named Entities: Structure and Translation. A Study Based on a Parallel Corpus (Arabic-Spanish-English)
See Named Entities.doc

Tam s V radi: Taking stock of the Bilingual Lexicon
See Taking Stock of the Bilingual Lexicon.doc

Discourse

Nadine Aldinger: Corpus-driven genitive disambiguation

See Corpus-driven gesture disambiguation.doc

Minhee Bang: Representation of foreign countries in two US newspapers: premodifications of keywords, countries, country, nations and nation
See Representation of foreign countries in two US newspapers.doc

Michael Barlow: Input grammars and output grammars: Investigating the language of individual speakers Christian Chiarcos & Olga Krasavina: Rhetorical Distance Revisited: A pilot study
See Rhetorical Distance Revisited.doc

Huaqing Hong: SCORE: A Multimodal Corpus Database of Education Discourse in Singapore Schools
See Scope.pdf

Henk Louw: Really Too Very Much: Adverbial Intensifiers in Black South African English
See REALLY TOO VERY MUCH.doc

Ling Yin & Richard Power: Investigation of the structure of topic expressions: a corpus-based approach
See Investigation of the Structure of Topic Expressions.doc

Massimo Poesio & Ron Artstein: Annotating (anaphoric) ambiguity
See Annotating (Anaphoric) Ambiguity.pdf

Evaluation and Stance

Monika A. Bednarek: "He's nice but Tim" -- contrastive evaluation in the British press
See 'He's nice but Tim': contrast in British newspaper discourse.doc

Sara Radighieri: Arts in the news: Evaluative language use in the 'arts review'
See Arts in the news.doc

Grammar

Solveig Granath & Michael Wherrity: Prepositions with that-clause complements in tagged corpora, with a special focus on in that
See Prepositions with that-clause complements in tagged corpora.doc

Vladimir Petkevic & Frantisek Cermak:Linguistically motivated tagging as the base for a corpus-based grammar
See Linguistically Motivated Tagging as a Base for a Corpus-Based Grammar.doc

Simone Sarmento: Distribution of Modal Verbs in an Aviation Corpus
See Distribution of Modal Verbs in an Aviation Corpus.doc

Chris Shei: Analysing Chinese Sentence-final Particles Using Academia Sinica Balanced Corpus of Modern Chinese
See Analysing Chinese Sentence.doc

Seo-in Shin: Automatic Pattern Extraction for Korean Sentence Parsing
See Automatic Pattern Extraction for Korean Sentence Parsing.doc

Language Learning & Error Analysis through Corpora

Mariko Abe and Yukio Tono: Variations in L2 spoken and written English: investigating patterns of grammatical errors a cross proficiency levels
See Variations in L2 spoken and written English.doc

Mar a Bel n D ez Bedmar-Struggling with English at University level: error patterns and problematic areas of first-year students interlanguage
See Bedmar Uni English.doc

Xiaotian Guo: Modal Auxiliaries in Phraseology: A Contrastive Study of learner English and NS English

See A Contrastive Study of Learner English and NS English.doc

Anke L deling, Peter Adolphs, Emil Kroymann & Maik Walter: Multi-level error annotation in learner corpora
See Multi-level error annotation in learner corpora.doc

Zhang Yang: College English Course Corpus

Language Processing & Corpus Tool

Sabine Bartsch, Elke Teich, Monica Holtz & Richard Eckart: Corpus-based register profiling of texts from mechanical engineering
See Corpus-based register profiling of texts.pdf

Anja Belz: Corpus-driven Generation of weather Forecasts
See Corpus-driven Generation of weather Forecasts.pdf

Pernilla Danielsson & Andrew Sayers: Enhancing Concordance Method: Introducing the CHAB

Stefan Evert & Manuela Schonenberger : Separating the sheep from the goats: Clarifying corpus content using XML
See Separating the sheep from the goats.doc

David Hardcastle: Using the distributional hypothesis to derived co-occurrence scores from the British National Corpus
See Using the distributional hypothesis.doc

Laura L fberg Scott Piao, Asko Nykanen, Krista Varantola, Paul Rayson and Jukka-Pekka Juntunen: A semantic tagger for the Finnish language
See A semantic tagger for the Finnish language.doc

Yuji Matsumoto, Masayuki Asahara, Kou Kawabe, Yurika Takashi, Yukio Tono, Akira Ohtani and Toshio Morita: ChaKi: An Annotated Corpora Management and Search System

See ChaKi.doc

D bora Oliveira, Diana Santos, Luis Sarmento & Belinda Maia: Corpus analysis for indexing: when corpus-based terminology makes a difference
See Corpus analysis for indexing.doc

Shih-Ping Wang: Integrating corpora and word-focused tasks into a linguistics project for word growth
See Integrating corpora and word-focused tasks into a linguistics project.doc

Maria ZIMINA- Bi-text topography and quantitative approaches of parallel text processing
See Bi-text Topography and Quantitative Approaches.doc

Eros Zanchetta and Marco Baroni: Morph-it! A free corpus-based morphological resource for the Italian language
See Morph-it!.doc

The Lexicon

Antti Arppe: The role of morphological features in distinguishing semantically similar words
See The role of morphological features in distinguishing semantically similar words.doc

J rg Asmussen: Automatic determination of new words within domain-specific vocabularies using document classification and frequency profiling
See Automatic detection of new domain-specific words.

Marco Baroni & Stefan Evert: Testing the extrapolation quality of word frequency models
See Testing the extrapolation quality.pdf

Dr Paul Doyle: Replicating Corpus-Based Linguistics: Investigating Lexical Networks in Text
See Replication and Corpus Linguistics.pdf

Cvetana Krstev & Dusko Vitas : Corpus and Lexicon Mutual In-completeness
See Corpus and Lexicon.doc

Jennifer Pedler: Using semantic associations for the detection of real-word spelling errors
See Using semantic associations for the detection of real-word spelling errors.doc

Scott S.L. Piao, Dawn Archer, Olga Mudraya, Paul Rayson, Roger Garside, Tony McEnery, Andrew Wilson: A Large Semantic Lexicon for Corpus Annotation
See A Large Semantic Lexicon for Corpus Annotation.doc

Elisabete Marques Ranchhod: Using Corpora to Increase Portuguese MWE Dictionaries. Tagging MWE in a Portuguese Corpus.
See Using Corpora to Increase Portuguese MWE Dictionaries.pdf

Sofie Van Gijsel, Dirk Speelman & Dirk Geeraerts: A Variationist, Corpus Linguistic Analysis of Lexical Richness
See Lexical Richness.doc

Phraseology & Patterns in language

Frantisek Cermak & Michal Křen: Large Corpora, Lexical Frequencies and Coverage of Texts
See Large Corpora, Lexical Frequencies and Coverage of Texts.doc

Christopher Gledhill & Pierre Frath: A Reference-based Theory of Phraseological Units: the Evidence of Fossils.
See A Reference-based Theory of Phraseological Units.doc

Eva Hajičov , Jiri Havelka & Katerina Vesela: Corpus Evidence of Contextual Boundness and Focus
See Corpus Evidence of Contextual Boundness and Focus.doc

Csaba Oravecz, Karoly Varasdi & Viktor Nagy: Lexical idiosyncrasy in MWE extraction
See Lexical idiosyncrasy in MWE extraction.doc

Bertus van Rooy: Expressions of modality in Black South African English
See Expressions of modality in Black South African English.doc

Petra Storjohann: Corpus-driven vs. corpus-based approach to the study of relational patterns
See Corpus-driven vs. corpus-based approach.doc

Christiane Wanzeck: The Determination of Phraseological Units in Historical Corpora: An Analysis System for Early New High German
See The Determination of Phraseological Units in Historical Corpora.doc

The Web as a corpus

Abdulrahman Almuhareb & Massimo Poesio: Finding Attributes in the Web
See Finding Attributes in the Web Using a Parser.pdf

Ilias Koutsis, Geroge Kouklakis, George Mikros & George Markopoulos: MINOTAVROS A tool for the semiautomated creation of large corpora from the Web.
See Minotavros.doc

Alexander Mehler & Rudiger Gleim: Polymorphism in Generic Web Units A Corpus

Linguistic Study (CLC)
See Alexander_Mehler_and_Ruediger_Gleim_Corpus_Linguistics_2005.pdf

Antoinette Renouf: The WebCorp Search Engine: a holistic approach to web text search
See The WebCorp Search Engine.doc

Jes s Tom s, Francisco Casacuberta & Jaime Lloret: WebMining: Non supervised system to obtain parallel corpus from the Web
See WebMining.pdf

Motoko Ueyama & Marco Baroni: Automated construction and evaluation of a Japanese web-based reference corpus
See Automated Construction and Evaluation of Japanese Web-based Reference Corpora.doc

Spoken Discourse

Adriano Allora: A Tentative Typology of Net-mediated Communication
See A Tentative Typology of Net-mediated Communication.pdf

Knut Hofland & Annette Myre Jorgensen: COLA: A Spanish spoken corpus of youth language
See COLA.doc

Kikuo Maekawa: Quantitative Analysis of Word-form Variation Using a Spontaneous Speech Corpus
See Quantitative Analysis of Word-form Variation.doc

Antonio Moreno-Sandoval & Ana Gonzales-Ledesma: Pragmatic analysis of man-machine interactions in a spontaneous speech corpus
See Pragmatic analysis of man-machine interactions.doc

University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
Tel: +44 (0)121 414 3344