Meta
Journal des traducteurs
Translators' Journal

META

Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus)

Cristina Monti, Claudio Bendazzoli, Annalisa Sandrelli and Mariachiara Russo

Volume 50, Number 4, décembre 2005

Pour une traductologie proactive — Actes
For a Proactive Translatology — Proceedings
Por una traductología proactiva — Actas

URI: https://id.erudit.org/iderudit/019850ar
DOI: https://doi.org/10.7202/019850ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

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). https://doi.org/10.7202/019850ar

Article abstract

Parallel corpora have long been awaited in simultaneous interpreting studies in order to validate existing theories and models. The present paper illustrates the development of the European Parliament Interpreting Corpus (EPIC), an open, parallel, multilingual (English, Italian and Spanish), POS-tagged corpus of European Parliament source speeches and simultaneously-interpreted target speeches. The aim of the project is to study recurrent lexical patterns and morphosyntactical structures across all the possible language combinations and directions, and verify empirically whether different strategies can be detected when interpreting from a Germanic language into a Romance one and vice-versa, or between two Romance languages. EPIC is freely available on-line for the research community to use and contribute to.

Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus)

CRISTINA MONTI, CLAUDIO BENDAZZOLI, ANNALISA SANDRELLI, MARIACHIARA RUSSO[1]
University of Bologna, Forlì, Italy
cbendazzoli@yahoo.com

RÉSUMÉ
Les corpus parallèles dans le domaine de la recherche sur l'interprétation simultanée étaient attendus depuis longtemps pour valider des théories et des modèles existants. La présente contribution a pour but de présenter EPIC (European Parliament Interpreting Corpus), un corpus ouvert, parallèle, multilingue (anglais, italien et espagnol) et avec étiquetage des parties du discours, composé de discours source prononcés au Parlement européen et de discours cible interprétés en simultanée. Le but de ce projet est d'examiner les modèles lexicaux et les structures morpho-syntaxiques dans toutes les combinaisons linguistiques considérées et quelles que soient la langue de départ et d'arrivée, et de vérifier de manière empirique si des stratégies différentes peuvent être décelées lors d'une interprétation à partir d'une langue germanique vers une langue romane et vice-versa, ou entre deux langues romanes. EPIC est librement accessible en ligne pour les chercheurs et est ouvert à leurs contributions.

ABSTRACT
Parallel corpora have long been awaited in simultaneous interpreting studies in order to validate existing theories and models. The present paper illustrates the development of the European Parliament Interpreting Corpus (EPIC), an open, parallel, multilingual (English, Italian and Spanish), POS-tagged corpus of European Parliament source speeches and simultaneously-interpreted target speeches. The aim of the project is to study recurrent lexical patterns and morpho-syntactical structures across all the possible language combinations and directions, and verify empirically whether different strategies can be detected when interpreting from a Germanic language into a Romance one and vice-versa, or between two Romance languages. EPIC is freely available on-line for the research community to use and contribute to.

MOTS-CLÉS/KEYWORDS
directionality, parallel corpora, simultaneous interpreting, EPIC, electronic corpus

Introduction

The spectrum of interpreter-mediated events is wide-ranging and so is the array of research paradigms and methodologies adopted so far. The body of knowledge on interpreting has been increasing steadily, especially since the '80s (Pöchhacker 1995), through observational and experimental studies, based on case-studies or limited samples of data, which have produced insightful results. The hypotheses suggested and the explanatory models developed, however, need to be validated or refuted on large and homogeneous sets of data. But, as has already been pointed out by Gile (1994, 1997, 2000), Kalina (1994) and Shlesinger (1998b), among others, a number of practical and methodological hurdles beset interpreting research. A significant objective obstacle is that interpreting data are not easily available and accessible. Gathering authentic interpreting data from conferences is a daunting task. Indeed, recordings of original speeches (source text, henceforth ST) are difficult to obtain because consent may be denied by organisers and speakers for reasons of confidentiality or lack of understanding towards scholarly and teaching purposes.

Similarly, recordings of interpreting performances (target texts, henceforth TT) are rarely made available by interpreters, who fear other colleagues' judgements. Furthermore, besides being limited in number, these recordings may not be of the quality required to carry out adequate analyses. With regards to the methodological issues affecting interpreting research, lack of homogeneity in interpreting data and research designs and lack of consistency between the object of study and the means of investigating it are only some of the obstacles, which do not allow for reliable comparisons of results or the exchange of material among researchers. This prevents researchers from validating the interesting trends observed on larger interpreting samples. Against this backdrop, the introduction of a corpus-based approach to the study of interpreting, as already advocated by Shlesinger (1998a) who called for the creation of parallel interpreting corpora, would certainly mark a turning point in this field.

The present paper presents the electronic corpus EPIC (European Parliament Interpreting Corpus) we are currently creating to study the effects of directionality in simultaneous interpreting (henceforth, SI). We use the term **directionality** with two meanings, firstly to refer simply to language combination and direction (e.g. English into Italian vs. Italian into English) and, secondly, to compare and contrast interpreting into the mother tongue ("A" language) vs. interpreting into a foreign language ("B" language). The latter was a taboo in EU institutions until the latest stage of enlargement but, at the same time, has always been common practice in the Italian conference market. Interpreting into B, or *retour*, has become frequent and worth investigating also for its pedagogical implications, as is shown by recent publications on the topic (Falbo et al. 1999, Kelly et al. 2003, Donovan 2004).

Our present objective is to investigate interpreters' strategies, recurrent lexical patterns and morpho-syntactical structures depending on the language combinations *per se*. In particular, our project concerns English, Italian and Spanish in all the possible combinations and directions: we wish to verify whether, and to what extent, a noticeable effect can be observed when interpreting from a Germanic language into a Romance one and vice-versa, or between two Romance languages. To arrive at meaningful findings, we need large quantities of homogeneous data to be analysed electronically, not just collections, however large, of interpreted speeches stored in a computer for manual analysis, as has hitherto been the case. Hence, the "epic" effort of our interdisciplinary research group set up a year and a half ago to create a multimedia digital archive of interpreted and original oral texts (speeches) which are being transcribed, tagged and lemmatised to develop an electronic corpus. EPIC, which at present consists of over 172.000 words,[2] is an open, parallel and multilingual (English, Spanish and Italian) corpus available on-line for the whole interpreting community to contribute to and share (see Web references), so as to advance our knowledge and understanding of interpreting and further enhance its teaching.

In the first section the rationale and methodology followed to create EPIC are explained, focussing on data characteristics, collection and organisation. Transcription criteria and procedures are described in the second section, while the procedures for POS-tagging, lemmatising, indexing and querying the transcripts make up the third and final section.

1. The EPIC multimedia archive

The first task of our research group was the creation of a digital multimedia archive containing both STs and TTs.

1.1 Material selection

The selection of material is crucial in determining the representativeness of the corpus (Halverson 1998, Bowker and Pearson 2002). As was highlighted in the Introduction, SI studies are hampered by practical and methodological obstacles. Firstly, collecting a large amount of high quality interpreting data is problematic. Secondly, from a methodological point of view, it must be highlighted that "to establish ecological validity and to arrive at meaningful findings, one must control as many of the independent variables as possible, so as to ensure that measurements in terms of the chosen dependent variable(s) are indeed reliable indicators of whatever one wishes to

measure" (Shlesinger 1998b: 3-4). In our case, several variables including, among others, the interpreters, their working conditions, the setting, the speakers and the topics discussed must be kept under control in order to have homogeneous and reliable data through which directionality can be studied.

In the light of the above, the **plenary sittings**[3] of the European Parliament were chosen as source material because they provide a solution to a number of problems. Starting with quantity and availability, EP part-sessions are held monthly and simultaneous interpretation is provided in all the EU official languages. A large portion of all the part-sessions is broadcast live by the satellite TV channel EbS (Europe by Satellite) which enables viewers to select the language channel. Permission to use the material for educational and research purposes was duly obtained from the EP Audiovisual Archive and EbS.

Another reason for choosing this material is its degree of homogeneity, in that all speeches are produced in the same highly formal and institutionalised setting (de Manuel Jerez 2003a, 2003b, Marzocchi and Zucchetto 1997), and the interpreters are all qualified and experienced professionals who have passed a strict selection procedure and usually work into their mother tongue.

This material also offered other advantages: in particular, the EP website publishes the verbatim reports of the debates and provides a wealth of information about the speakers and the topics they discussed (see 2.4).

1.2 Data collection and organisation

Data collection and organisation have been a very challenging part of the project. A high degree of co-ordination among data collectors was necessary, together with the establishment of a set of steps and standardised procedures, namely data collection, digitisation, file editing, archiving and cataloguing.

The first step was recording the material using four workstations comprising TV sets, videotape recorders and satellite decoders at the Department of Interdisciplinary Studies in Translation, Languages and Cultures (SITLeC) and at the Advanced School for Interpreters and Translators (SSLMIT) of the University of Bologna at Forlì.[4] The four workstations were set to the following language channels: 1) original debate 2) English 3) Spanish 4) Italian. This way, the research group obtained original speeches in the three languages and the two corresponding interpreted versions for each source speech, therefore allowing for comparisons among different language directions.

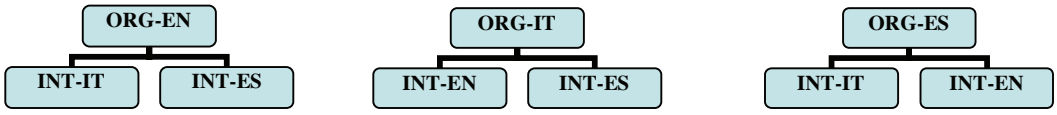

Fig. 1. Corpus architecture (ORG = original speech, i.e. source text; INT = interpreted speech, i.e. target text; EN = English; IT = Italian; ES = Spanish)

Overall, five part-sessions were recorded in 2004.[5] The original language recordings contain speeches in all EU official languages. Owing to the EbS schedule, the collected data include not only parts of the plenary sessions, but also interviews, press conferences and reports on EU topics. All this extra material has been kept for possible future research projects and/or as teaching material.

Digitising video-recordings means converting analogue data into digital format. Several software products available on the market were taken into account, and a choice had to be made regarding formats which had to be flexible and in line with the needs of our study and future exchange of materials with other researchers. To optimise file size, original speeches were digitised as .mpeg video-files with the software *Pinnacle Studio*. Interpreted speeches were saved as .wav audio files using *WaveLab* and *CoolEdit*.[6] The process of digitisation is still going on. At the time of writing, 91 videotapes have already been digitised.

Digitised data are then edited and catalogued in the EPIC multimedia archive. Editing consists in selecting the source speeches in the languages under study from the digital files of the EP sittings, together with their corresponding interpreted versions, in order to create a set of video and audio clips. In order to manage data efficiently, a classification system with a reference code assigned to each file has been devised (see 3.1). At present, over 250 video clips have been produced in this way, together with 500 audio clips.

The EPIC multimedia archive is stored on a dedicated external hard-disk. It contains the digital files of the plenary debates (unedited, in all the EU languages), additional recorded material (press conferences, EU stock footage, and so on) and the edited clips, i.e. video files of the source speeches in the three languages, audio files of the corresponding interpreted versions, and the corresponding transcripts. Section 2 illustrates the transcription process.

2. Transcribing the material

Once the video and audio material is ready, it must be transcribed. Transcribing is generally considered a time-consuming and labour-intensive activity. Moreover, in the field of SI studies there is a lack of standard conventions for annotating speech and interpreting features. In fact, too many different conventions are employed by researchers, which poses several problems to sharing and exchanging data (Cencini 2002).

A preliminary assumption of the present study is that transcripts represent a first level of analysis in themselves, in that transcription is by its very nature a selective process: "The transformation of audio and/or video recordings into a written format, a transcript, represents yet another selection process. That is, the transcription system used and the variations in individual transcribers' practices introduce directly and specifically the analysts' interests and theories" (Psathas and Anderson 1990: 75).

Indeed, it is virtually impossible to reproduce all the characteristics of speech in writing, as there are several levels (i.e. linguistic, paralinguistic and extra-linguistic) comprising an infinite number of features, such as pauses, repetitions, prosody, body language, and many more. The specific type of material under investigation and the aim of one's research are among the most significant factors that influence the way oral texts are transcribed for later analysis. As described in the previous sections, the present research project aims at analysing a large amount of original and simultaneously interpreted speeches through corpus linguistics techniques. To streamline the procedure, we concentrated first on a basic level of annotation, which would serve then as a starting point for further levels (Armstrong 1997: 158).

In order to establish what should be annotated and how, a review of transcription systems and conventions published so far was carried out. The notation systems used in other SI studies were considered, as well as conventions employed in undergraduate dissertations by SSLMIT students in Forlì.

We reached the conclusion that the Jeffersonian system was best suited to be the main reference system for our purposes, in that it is well-established and widely accepted in the research community,[7] as can be seen in many studies on conversation analysis and interpreting (Orletti and Testa 1991, O'Connell and Kowal 1994, Straniero Sergio 1999).

A first set of selected features included vowel and consonant lengthening, latching, sighs, mispronounced words, truncations and pauses, while punctuation was used as prosodic marker. However, this method soon proved impractical in view of preparing a large amount of **machine-readable data**. Therefore, a basic annotation set was chosen for each transcription level (i.e. linguistic, paralinguistic and extra-linguistic levels), as described in the following subsections.

2.1 Linguistic level

All the words uttered by both speakers and simultaneous interpreters are transcribed orthographically. There are no punctuation signs in the transcripts, as they are typical of written texts, they may not be suitable for automatic analysis and there is no established correspondence between the duration of pauses in speech and the various punctuation signs. Transcribed texts are

segmented in units of meaning, on the basis of the speaker's intonation and syntactic information in the sentence involved. The double bar sign // is used to indicate the end of each segment. This segmentation is mainly functional to the alignment between source and target texts, a future step in our project.[8]

Spelling conventions follow the standards applied in EU official documents. These indications can be found in the Interinstitutional Style Guide available on the European Parliament website for all the official languages of the Union (see Web references). Numbers, dates and percentages are fully spelt out.

2.2 Paralinguistic level

In the present study annotations at this level are limited to truncated and mispronounced words. In order to perform POS-tagging and automatic analysis (see section 3.2), mispronounced words and those that are improperly articulated by speakers or interpreters must be spelt correctly, so that the computer can recognise them and process them (Leech et al. 1995). Thus, such words are "normalised" first; each normalised item is then followed by the word as it was actually uttered in angular brackets < >. Depending on the kind of analysis to be carried out, the words in brackets can be included or excluded automatically from the corpus (see 3.3).

For truncated words (i.e. words that are not fully uttered) the - symbol is attached to the end of the word (e.g. "Pre- President it is a ple- pleasure to be here"), while for words featuring an "internal truncation" (i.e. words that are fully uttered but with interruptions in the speaker's articulation) the _ symbol is used to link the two word chunks, preceded by the normalized version (e.g. "this is important </im_portant/> for all the countries"). Mispronounced words are enclosed between bars (e.g. "cholera </chorela/>").

Pauses are also included, but they are currently annotated on the basis of the transcriber's perception only. Both silent (...) and filled (ehm) pauses are considered, though no details are provided about their duration. This is an attempt to make oral data reflect the mode of delivery as close as possible, while preserving readability, i.e. to obtain a written representation of speech by means of user-friendly transcripts. Moreover, this constitutes a basis for future systematic pause annotation to be carried out using appropriate electronic tools, thus providing exact information about pause duration and their location.

2.3 Extra-linguistic level

This level provides information about the transcript file (e.g. date, language, etc.), the speaker (e.g. name, gender, political function, country of origin, etc.) and the speech itself (e.g. number of words, type of delivery, speed, topic, etc.). All this information is presented in a **header** containing a number of fields. These come before the transcribed text and were used to set the parameters to carry out automatic queries (see section 3.1).

The EPIC transcription conventions are summarised in the following table:

SPEECH FEATURE	EXAMPLE	TRANSCRIPTION CONVENTION
Word truncations	propo pro posal	propo- proposal </pro_posal/>
Mispronounciations	chorela	cholera </chorela/>
Pauses	(filled / empty)	ehm …
Numbers	532	five hundred and thirty-two

Figures Dates	4% 1997	four per cent nineteen ninety-nine
Unintelligible		#
Units	(based on syntax and speaker's intonation)	//

Table 1: EPIC transcription conventions.

2.4 Transcription process

Against this background, efforts were made to ease and speed up the transcription process. The very nature of the material under study, i.e. European Parliament speeches, offered some advantages in this respect. It was possible to establish a "transcription procedure" consisting in two main stages: producing a draft transcript very quickly in the first place, and then producing a final draft to be used for analysis.

As regards producing the preliminary draft, a distinction has to be made between source and target speeches.

The ST transcripts are easily obtained from the EP verbatim report, which is made available on the EP website a short time after each part-session. Obviously, the texts in the verbatim report do not reflect speech features very closely, as they undergo stylistic revision, punctuation is added and speakers' mistakes are amended (e.g. there are no instances of unfinished sentences, mispronounced words and ungrammatical structures, to name just a few). However, this written material provides a very useful basis to obtain the final draft transcript, in which speech features are closely reproduced.

As regards the TTs (simultaneous interpretations), these are not transcribed by EP officials. The verbatim report available in all EU official languages is the result of a written translation, and no reference is made to the interpreters' renderings. This means that all the TTs have to be transcribed from scratch. Since in our research group we are all trained conference interpreters, we use speech recognition programs whilst performing **shadowing** (Schweda Nicholson 1990, Lambert 1992), i.e. listening to an oral text and simultaneously repeating it aloud. This way, transcribers listen to the interpreter's version and repeat it aloud using a microphone connected to computers which have been trained to recognise their voices.[9] Thus, a first draft transcript is "automatically" generated. The second stage involves revising first draft transcripts and adding speech feature annotations included in the study.

Finally, transcripts are cross-checked, so as to minimise mistakes and reduce the effects of individual transcribers' perception abilities. Once cross-checking is finished, the transcripts are saved in text format in the archive and are ready to be tagged and processed for automatic analysis (see 3).

As a general rule, much care was taken in choosing basic ASCII characters, in order to avoid computer-readability problems (Leech et al. 1995). For the same reason, extreme care had to be given to avoiding improper or unintentional use of bar spaces in transcript files. If extra spaces do not pose any problem to manual analysis, the same cannot be said of automatic analysis, in which even a single typing mistake can stop the whole system from functioning properly.

Indeed, the two guiding principles followed in the EPIC transcription process are **machine-readability** and **user/annotator-friendliness**. In other words, transcripts were designed to be easily produced and easily accessible for both manual and automatic analysis.

3. From manual to automatic analysis

3.1 The header

As was mentioned in 2, the transcripts contain extra-linguistic information about each speech in the form of a **header** which precedes the text. The information recorded in the header fields is then

used to query the corpus through a dedicated Web interface, as is described in 3.3. The following is an example of the template used:

(date: 25-02-04-p
speech number: 033
language: it
type: org-it

duration: short
timing: 85

text length: short
number of words: 153

speed: low
words per minute: 108

source text delivery: impromptu

speaker: Fatuzzo, Carlo
gender: M
country: Italy
mother tongue: yes

political function: MEP
political group: PPE-DE

topic: Politics
specific topic: Annual Policy Strategy of the European Commission for 2005

comments: NA)

The first four fields in the header make up a reference code used to classify the speeches. The first one is the date on which the speech was delivered (day, month and year), followed by a letter which indicates whether it was a morning or an afternoon sitting ("m" or "p", respectively).[10] Then, a progressive number is assigned to all the transcripts for easy retrieval of each speech. The reference code is completed by two fields with information on the language ("en", "it" and "es" for a speech in English, Italian or Spanish) and the type of speech (an original speech vs. an interpretation, i.e. "org" or "int"). The above example was an Italian source speech delivered on 25 February 2004, during the afternoon sitting.

The next group of fields contains information on some speech features: speech duration, text length, speed and mode of delivery. Since our aim is to use these data to carry out targeted queries in the corpus, the information must be easy to process automatically. Therefore, as well as recording the exact figures indicating the number of seconds, the number of words and the words per minute in the corresponding fields, a label had to be assigned to classify speeches as short, medium or long (in terms of text length and duration) and as speeches delivered at a low, medium or high speed. The values covered by each label were established on the basis of the present corpus of speeches, and therefore can only be considered representative of this material, namely the speeches delivered during a specific series of plenary part-sessions of the European Parliament. In particular, we have established the following reference values for the EPIC corpus:

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 w/m

Clearly, in different contexts these values do not apply, e.g. in the Italian conference interpreting market a speech delivered at 150 w/m is fast, not medium. However, the European Parliament has established very strict rules for the allocation of speaking time to MEPs, who generally speak very fast in an attempt to say as much as possible. Therefore, in this sense and in this particular context, 150 w/m can be considered average. Another factor influencing speed is the mode of delivery, namely, whether speakers read out from a script, improvise or alternate between the two modes. This information can be easily gleaned from the video clips, and therefore is included in a dedicated header field.[11]

The topic of the speech is classified on the basis of macro-categories, such as economics and finance, politics, etc. Information on the specific topic under discussion is also provided in a separate field, which contains the exact heading used by EP officials in the verbatim reports.[12] The header is completed by a group of fields with the following information about the speaker: name, gender, country of origin, mother tongue, political function and political group. When the speaker is an interpreter, no values are assigned to the fields "name", "country", "political function" and "political group", because this type of information is either not known or not applicable. When the speaker is a source language speaker, however, it is important to classify his/her political function. The speeches in the EPIC corpus are delivered by several types of speakers, classified as follows: MEPs, President of the European Parliament, Vice-President of the EP, representative of the European Commission, representative of the European Council of Ministers, or guest visiting the Parliament. If the speaker is an MEP, we also indicate the political group. Likewise, if the speaker is a Commissioner or a European Council Minister, we indicate their political responsibilities (i.e. the field of action of the Commissioner or the European Council configuration) in the last header field which we also use for comments, for example to indicate a speaker's non-standard accent, a technical problem in the recording, or any other unusual feature considered potentially important for later analysis.

All of the fields described above are used to set the search filters of the EPIC web interface, which is described in 3.3. Before the corpus can be queried, however, it needs to be POS-tagged, lemmatised and indexed.

3.2 POS-tagging EPIC

POS-tagging means assigning a part-of-speech label (**tag**) to each word in a corpus, in order to make it possible to search it automatically for specific patterns and structures. In our case, when the EPIC corpus is fully tagged, the main focus will be on comparative studies of patterns across the different language combinations and directions described in 1.2 (see figure 1).

POS-tagging can be done automatically by using dedicated software programmes called **taggers**. The main stages of the tagging process are the following: **tokenization**, **tag assignment** and **disambiguation** (Bowker and Pearson 2002: 84). Tokenization means breaking down the text into individual words and punctuation signs. Then, the tagger assigns a part-of-speech tag to each **token** (item in the corpus), using various morphological and context-based cues to decide the right tag for ambiguous words. Different taggers take such decisions by using different methods. In particular "stochastic taggers 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 and Martin 2004: 17-18). In other words, **stochastic taggers** involve manually-annotating a training corpus, which is then used by the tagging algorithm to extract generalisations. Since the training corpus is necessarily small, when a tagger encounters a previously unseen word, it applies the rules it has extracted to perform probability calculations and assign the most likely tag. Clearly, for automatic tagging to produce accurate results, the texts in the corpus to be tagged must be as similar as possible to the texts in the training corpus.

The (stochastic) taggers chosen for EPIC are *TreeTagger* (Schmid 1994) for English, *Freeling* (Carreras et al. 2004) for Spanish, and the combination of taggers described by Baroni et al. (2004) for Italian.[13] It must be noted that although the accuracy rate of the chosen taggers is generally very high, they were developed for written texts, not for transcripts of spoken texts. As was explained in 2, we have greatly limited the number of speech features included in our transcripts. Nevertheless, the STs in EPIC display many features of spoken language, whereas the TTs are simultaneously interpreted texts with very special characteristics.[14] Therefore, in order to further improve the efficiency of the taggers on our data, our next step will be to manually correct any errors in a subset of English, Italian and Spanish texts (that is, to create our training corpora), and then feed them to the taggers for them to "update" their rules.

In order to be able to query the corpus, the tagged output is converted into XML format, and indexed by using the *IMS Corpus Work Bench – CWB* (Christ 1994). **Positional attributes** are thus associated to all individual words, in order to easily retrieve all the occurrences of each word in the corpus; the header fields in each transcript (see 3.1, above) are used to set the XML **structural attributes** which allow us to restrict queries on the basis of speech or speaker features, as in the example below.

```
<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>
```

A user-friendly web interface has been developed to make corpus querying simpler and faster. This is described in 3.3, below.

3.3 The EPIC web interface

The EPIC web interface is hosted by the SSLMIT Development web site, which includes several other corpus-based projects and useful resources for translators, interpreters and terminologists.[15] It features a number of information pages on the EPIC project, a simple query page, an advanced query option, and an interface to cwb-scan-corpus, which is a tool for the production of frequency lists.

At present, EPIC is made up of nine sub-corpora, namely three collections of STs in English, Italian and Spanish (org-en, org-it, org-es, respectively) and six collections of TTs in all the available language combinations and directions (int-en-it, int-en-es, int-it-en, int-it-es, int-es-en, and int-es-it). There are plans to align source texts and target texts for all the available language combinations and directions and then upload the resulting 6 aligned sub-corpora as well.

EPIC sub-corpora can only be queried separately. For example, if the aim is to compare English STs and TTs (i.e. the characteristics of original speeches produced in English and those interpreted into English from Italian and Spanish), separate queries must be issued in the English ST sub-corpus and in the English TT sub-corpora.

After selecting the desired sub-corpus, if users choose the **simple query** option, they can either interrogate the whole sub-corpus or restrict the search to a number of texts by using one or more of the **search filters** provided.

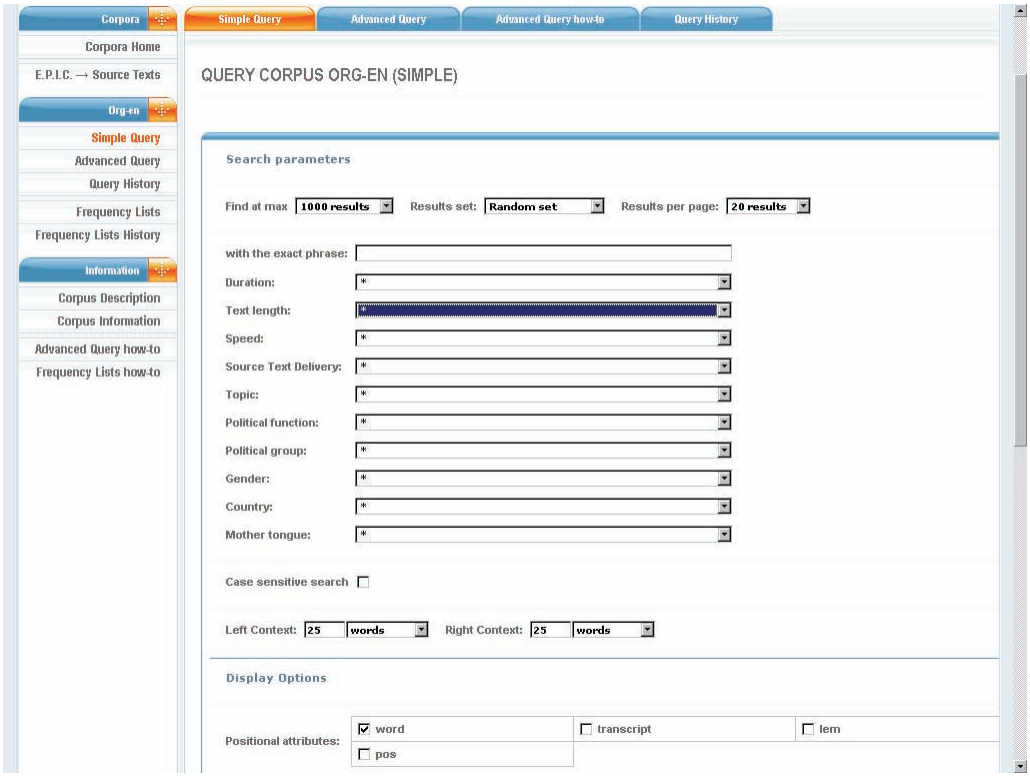

Figure 2: Simple query page

The "duration" search parameter makes it possible to search for a certain phrase in short, medium or long speeches (see 3.1). Similarly, the "text length" filter enables users to select speeches on the basis of the number of words in each speech, and the "speed" option to choose speeches delivered at low, medium or high speed.

The "source text delivery" option makes it possible to filter speeches according to delivery mode: read, impromptu, and mixed. The "topic" search parameter can be used to study the characteristics of EP speeches on agriculture and fisheries, or procedural matters, and so on. Users can also restrict their queries on the basis of speaker characteristics: political function (and political group in the European Parliament, where applicable), gender, country, and native language. The option "mother tongue" is particularly relevant for English, which is often used as a lingua franca by non-native speakers (e.g. Commissioners and Council Ministers often use English in the European Parliament).

The EPIC web interface enables users to issue **advanced queries** as well, by using the powerful CQP language of CWB. An information sheet with query hints and suggestions is available on the web site ("advanced query how-to"). 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 queries and advanced queries are visualized in a **KWIC (key-word-in-context) view**, with the queried word or string displayed in the middle of the screen and the specified left and right contexts (25 words by default). If a result seems interesting, it is also possible to look at the full text where that particular sentence can be found. The full text can be displayed in different ways: the XML attributes containing speaker and speech information (see figure 3), a "normalised" transcript with any disfluencies hidden (option "Show word"), the transcript reflecting how the words were actually uttered as closely as possible ("Show transcript"), the lemmatised transcript ("Show lem"), or the part-of-speech tags ("Show POS").

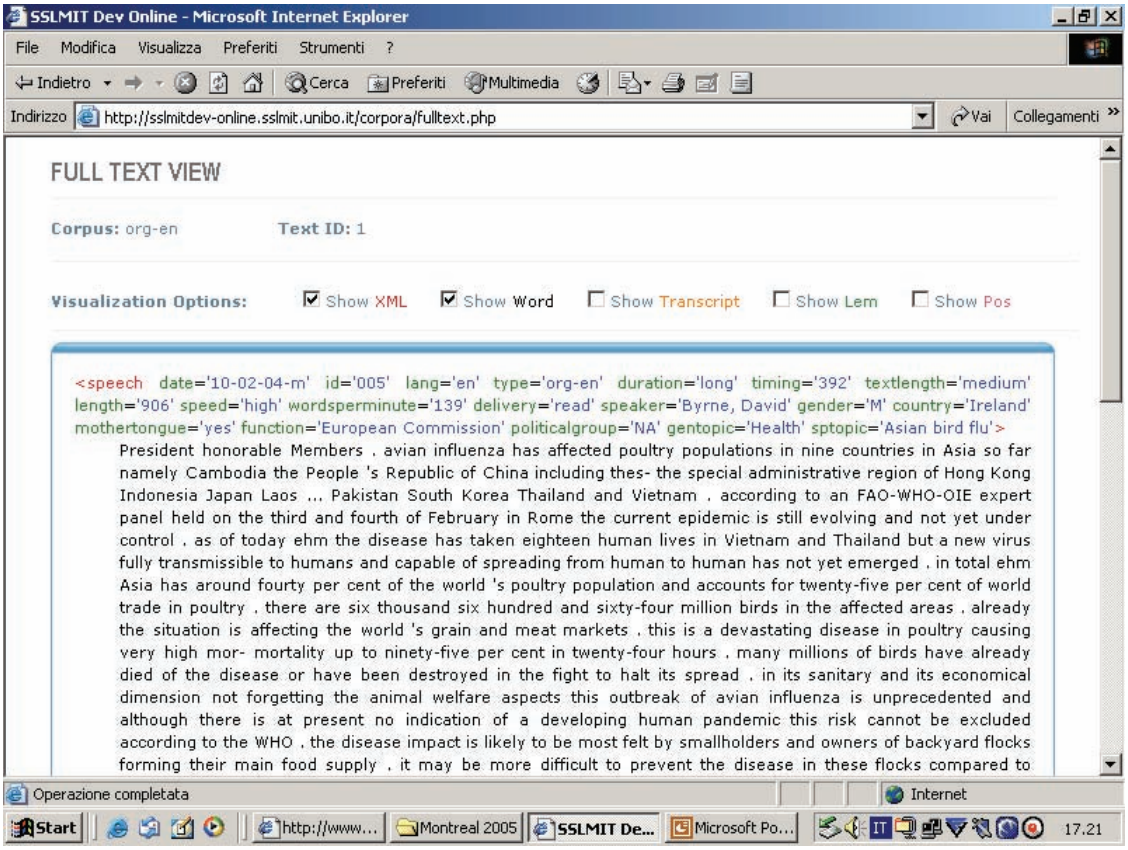

Figure 3: Transcript display options

Conclusions

Due to its nature, interpreting is probably the only field of study where it is possible to produce a parallel corpus of spoken language. The creation of an electronic parallel corpus is possible thanks to the collaboration of interpreters, computational linguists, corpus linguistics scholars and information technology experts. In our research group,[16] which reflects this multidisciplinary approach, great efforts have been made to harmonise complexity with simplicity, that is to say to combine corpus design and workflow organisation with easiness of use. All the steps taken in this research project have been extremely challenging and time-consuming, especially speech transcription and classification. Transcription conventions and procedures are designed to ensure fast production of machine-readable and user-friendly transcripts for both manual and automatic analysis. The basic transcription annotation adopted allow for several linguistic, paralinguistic and extra-linguistic analyses, but for more targeted queries, further levels of annotation could be added. This characteristic makes EPIC a flexible research tool.

With regards to data organisation and classification, we soon realised that in order to exploit our valuable resources to the full, we needed to include as much information as possible in the corpus: in other words, the wider and richer the corpus, the more detailed and articulated the material description and organisation. As was thoroughly illustrated in the previous sections, the EPIC web interface enables users to query this open, parallel, POS-tagged, lemmatised corpus by selecting the desired search parameters and to display corpus contents in different ways, thanks to structural and positional attributes. This opens up new research perspectives for interpreting.

Indeed, such a large collection of source and interpreted speeches lends itself to descriptive studies both on speakers' styles and speech genres (by selecting the search parameters "source text delivery" or "topic") and on interpreters' delivery norms and strategies. Another possible field of study could concern linguistic quality, by contrasting the language spoken in source

speeches (e.g. English produced by SL speakers) with the language of target speeches (e.g. English produced by interpreters). Equally insightful information could be collected by comparing morpho-syntactical structures and language patterns in relation to directionality (i.e. interpreting from or into a specific language) or to working language combination.

The research potential of EPIC parallels its pedagogical value. One of its manifold teaching applications could imply using the recorded material to expose trainees to EP interpreters' performances on a regular basis, so that they can learn to detect the distinctive features of professional quality. This could be particularly beneficial for trainees working into their B languages (a common practice in Italian academic institutions): since EP interpreters work into their A languages,[17] studying their performances when tackling specific interpreting problems will help trainees internalise adequate solutions and improve their skills. The multimedia archive is also a rich source of teaching materials (ST video files, TT audio files and transcripts) which could be exchanged with other universities, thus increasing the availability of training resources for interpreting and L2 learners. Furthermore, EPIC materials could be analysed for dissertations on SI, as is currently the case at the SSLMIT in Forlì.

As has been already stressed, EPIC is a parallel, open corpus which is constantly expanding as new material is being digitised and transcribed. At present, a sizeable part of EPIC contents is ready to be analysed. It is hoped that very soon it will be possible to discuss the first results obtained from investigating the corpus.

The next stage of the project envisages the development and deployment of an automatic text alignment procedure. In the future, we plan to further expand EPIC by adding a corpus of original speeches and interpreted versions produced by interpreters working into their B languages in order to study the other interesting aspect of directionality, i.e. *retour* interpreting.
Sharing resources benefits the scientific community and the advancement of the discipline. This is the spirit behind the choice of making EPIC an on-line resource: an interpreting research tool to share and contribute to.

Acknowledgments

We wish to thank Lorenzo Piccioni and Eros Zanchetta for creating EPIC web interface.

NOTES

1. Although the present article is the result of a joint effort, Mariachiara Russo can be identified as the author of the Introduction and Conclusions, Cristina Monti of section 1, Claudio Bendazzoli of section 2 and Annalisa Sandrelli of section 3.
2. This is an overall figure for all materials in all three languages. See section 1 for more details of the structure of EPIC.
3. In EP jargon, a **session** is a parliamentary year, a **part-session** is the EP monthly meeting, and a **sitting** is each of the daily meetings held during a part-session, as is explained in the relevant rule of procedure: http://www2.europarl.eu.int/omk/sipade2?PUBREF=-//EP//TEXT+RULES-EP+20040720+RULE-126+DOC+XML+V0//ENandHNAV=Y.
4. Obviously, it would have been better (and faster) to digitally record the broadcasts straight to computer, by using a digital decoder connected to our machines. Unfortunately, such technology was not available to us.
5. The recorded sessions are as follows: February, beginning and end of March (two part-sessions), April and July, with the newly-elected Parliament. 140 VHS tapes were used to this aim.
6. The chosen settings for the digital files are as follows: .mpeg (384 x 288 - 512 Kbits/sec - Freq. 44,1 Hz - Kbits/sec 64 bit); wav (Mono - 32.000 - 8 bit).
7. This system was first developed by Gail Jefferson and then adapted for a variety of research purposes, such as conversational analysis and interpreting studies.
8. Unlike content alignment, time alignment is currently beyond the scope of our research, but software tools, such as *Winpitch* and *Exmaralda* may prove useful in this respect.

9. The two programs used are *Dragon Naturally Speaking* and *IBM Via Voice*.
10. The letters stand for the Italian words for morning and afternoon (*mattino* and *pomeriggio*, respectively).
11. We have decided to include information on how the source language speech was delivered in the headers of the target language speeches as well, since interpreters may adopt different strategies according to the mode of delivery. For example, when speakers read out their texts, delivery tends to be fast, and interpreters may be forced to summarise or omit secondary information to keep up with the pace. Moreover, written texts have a lower redundancy and a higher information density, which is another complicating factor for the interpreter.
12. There is a wide range of topics in the corpus, reflecting the variety of debates in the EP.
13. Fairly obviously, in order to tag texts in different languages, different tagsets and rules must be used because of grammatical differences between languages.
14. In particular, filled pauses (transcribed as "ehm") and truncated words pose problems, as well as certain proper nouns, technical terms, loans and EU jargon.
15. The EPIC interface is based on the one created to query the *La Repubblica* corpus, a very large collection of articles published in one the main Italian dailies.
16. The other directionality group members are: Marco Baroni, Elio Ballardini, Silvia Bernardini, Gabriele Mack and Peter Mead.
17. As was briefly mentioned in the Introduction, the practice of having EP interpreters work into their A languages only has been slightly modified with the latest stage of EU enlargement. However, these changes do not generally apply to the English, Italian and Spanish booths, according to our recordings.

REFERENCES

ARMSTRONG, S. (1997): "Corpus Based Methods for NLP and Translation Studies", *Interpreting* 2-1/2, pp. 141-162.
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 Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R., 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, pp. 1771-1774
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 Oostdijk, Nelleke, Kristoffersen, Gjert and Geoffrey Sampson (eds.), *Compiling and Processing Spoken Language Corpora*, LREC 2004 Satellite Workshop, Fourth International Conference on Language Resources and Evaluation, 24 May 2004, pp. 33-39.
BOWKER, L. and J. PEARSON (2002): *Working with Specialized Language. A Practical Guide to Using Corpora*, London and New York, Routledge.
CARRERAS, X., CHAO I., PADRÓ, L. and M. PADRÓ (2004): "Freeling: An Open-source Suite of Language Analyzers", in Lino, Maria Teresa, Xavier, Maria Francisca, Ferreira, Fátima, Costa, Rute, and Raquel Silva (eds.), with the collaboration of Carla Pereira, Filipa Carvalho, Milene Lopes, Mónica Catarino and Sérgio Barros, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, ELRA, vol. 1, pp. 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, <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.
DONOVAN, C. (2004): "European Masters Project Group: Teaching Simultaneous Interpretation into a B language: Preliminary findings", *Interpreting*, 6-2, pp. 205-216.
FALBO, C., RUSSO, M. e F. STRANIERO SERGIO (a cura di) (1999): *Interpretazione simultanea e consecutiva*, Milano, Hoepli.
GILE, D. (1994): "Methodological Aspects of Interpretation and Translation Research", in Lambert, S. and B. Moser-Mercer (eds.), *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, Amsterdam-Philadelphia, John Benjamins, pp. 39-56.

GILE, D. (1997): "Interpretation Research: Realistic Expectations", in Klaudy, K. and J. Kohn (eds.), *Transferre necesse est*, Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting, 5-7 September 1996, Budapest, Hungary, Scholastica, pp. 43-51.
GILE, D. (2000): "Issues in Interdisciplinary Research into Conference Interpreting", in Englund Dimitrova, B. and K. Hyltenstam (eds.), *Language Processing and Simultaneous Interpreting: Interdisciplinary Perspectives*, Amsterdam-Philadelphia, John Benjamins, pp. 89-106.
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, pp. 494-513.
JURAFSKY, D. and J. H. MARTIN (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, <http://www.cs.colorado.edu/~martin/slp.html>.
KALINA, S. (1994): "Analysing Interpreters' Performance: Methods and Problems", in Dollerup, C. and A. Loddegaard (eds.), *Teaching Translation and Interpreting 2: Insights, Aims, Visions*, Amsterdam-Philadelphia, John Benjamins, pp. 225-232.
LAMBERT, S. (1992): "Shadowing", *The Interpreters' Newsletter* 4, pp. 15-24.
LEECH, G., MYERS, G. and J. THOMAS (eds.) (1995): *Spoken English on Computer: Transcription, Mark-up and Application*, New York, Longman.
MANUEL JEREZ, J. de (2003a): "El canal Ebs en la mejora de la calidad de la formación de intérpretes: estudio de un corpus en vídeo del Parlamento Europeo", in Collados Aís, Á., Fernández Sánchez, M.ª M. and D. Gile (eds.), *La evaluación de la calidad en interpretación: investigación*, Granada, Editorial Comares, pp. 207-218.
MANUEL JEREZ, J. de (2003b): "Nuevas tecnologías y selección de contenidos: la base de datos *Marius*", in Manuel Jerez, J. de (coord.), *Nuevas tecnologías y formación de intérpretes*, Granada, Editorial Atrio, pp. 21-61.
MARZOCCHI, C. and G. ZUCCHETTO (1997): "Some Considerations on Interpreting in an Institutional Context: The Case of the European Parliament", *Terminologie et Traduction* 3, pp. 70-85.
O'CONNELL, D. C. and S. KOWAL (1994): "Some Current Transcription Systems for Spoken Discourse: A Critical Analysis", *Pragmatics* 4, pp. 81-107.
ORLETTI, F. and R. TESTA (1991): "La trascrizione di un corpus di interlingua: aspetti teorici e metodologici", *Studi italiani di linguistica teorica e applicata* 20-2, pp. 243-283.
PÖCHHACKER, F. (1995): "Those who do, a profile of research(ers) in interpreting", *Target* 7-1, pp. 47-64.
PSATHAS, G. and T. ANDERSON (1990): "The 'Practices' of Transcription in Conversation Analysis", *Semiotica* 78-1/2, pp. 75-99.
SCHWEDA NICHOLSON, N. (1990): "The Role of Shadowing in Interpreter Training", *The Interpreters' Newsletter* 3, pp. 33-40.
SHLESINGER, M. (1998a): "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies", *META* 43-4, pp. 486-493.
SHLESINGER, M. (1998b): "Interpreting as a Cognitive Process: What Do We Know About How It Is Done?", Paper given at the *II Jornadas Internacionales de Traducción e Interpretación*, Málaga, 17-20 marzo 1997, Málaga, Grupo de Investigación de Lingüística Aplicada y Traducción de la Universidad de Málaga.
STRANIERO SERGIO, F. (1999): "The Interpreter on the Talk Show: Analyzing Interaction and Participation Framework", *The Translator* 5-2, pp. 303-326.

WEB REFERENCES

EbS (Europe by Satellite) TV channel: http://www.europa.eu.int/comm/dg10/ebs
EPIC interface on the SSLMITdev website:
http://sslmitdev-online.sslmit.unibo.it/corpora/corpora.php
European Parliament: http://www.europarl.eu.int
FreeLing: http://garraf.epsevg.upc.es/freeling/
IMS Corpus Workbench: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench
Interinstitutional style guide: http://publications.eu.int/code/en/en-000400.htm
SCHMIDT, T. (2001) "The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse". In Proceedings of the IRCS Workshop on Linguistic Databases, 219-227, in http://www.rrz.uni-hamburg.de/exmaralda/de/dokumentation.html

SCHMIDT, T. (2003a) *A short introduction to the EXMARaLDA Partitur-Editor,* in http://www.rrz.uni-hamburg.de/exmaralda/de/dokumentation.html
SCHMIDT, T. (2003b) "Visualising Linguistic Annotation as Interlinear Text". In *Arbeiten zur Mehrsprachigkeit*, Serie B (46) Hamburg, in http://www.rrz.uni-hamburg.de/exmaralda/de/dokumentation.html
TreeTagger: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/italian-tagset.txt