Claudio Bendazzoli & Annalisa Sandrelli (Forlì, Bologna)[1]

An Approach to Corpus-Based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus)

Contents

Abstract

Empirical research on simultaneous interpreting is hampered by the problem of collecting sufficient material (recordings of source speeches and interpreted target speeches) for the testing of hypotheses and validation of existing theories. In other words, corpora have long been awaited in the field of Interpreting Studies. In January 2004 a research group on corpus-based interpreting studies was set up in the Department of Interdisciplinary Studies in Translation, Languages and Cultures (SITLeC) of the University of Bologna at Forlì, in order to create an electronic parallel corpus with source and target speeches in Italian, English and Spanish. The main research interest of the group is the study of interpreters' strategies across different language directions (directionality) and language-pair related difficulties. In this paper, the authors illustrate the various stages of development of the EPIC corpus, highlighting both research and pedagogical applications of this "multidimensional" tool comprising video, audio and written materials, which can be retrieved, selected and analyzed by using corpus linguistics techniques.[2]

1 Introduction

In January 2004 a research group on corpus-based interpreting studies (the Directionality Research Group[3]) was set up in the Department of Interdisciplinary Studies in Translation, Languages and Cultures of the University of Bologna at Forlì, in order to study conference interpreters' strategies across different language directions (directionality) and language-pair

[1] Although the present article is the result of a joint effort, Claudio Bendazzoli can be identified as the author of (3), (4), and (5.1), whereas Annalisa Sandrelli is the author of (1), (2), and (5.2). The Conclusions (6) were jointly drafted.

[2] For an update on investigations on the basis of the EIPC corpus cf. Annalisa Sandrelli/Claudi Bendazzoli, The EPIC corpus - first results. Proceedings of the third Mutra Conference 'LSP Translation Scenarios'. Vienna 30th April - 4th May 2007. Edited by Sandra Nauert. Saarbrücken: www.euroconferences.info/proceedings.

[3] 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.

related aspects. The chosen method was the application of corpus analysis techniques to interpreting data. An electronic parallel corpus of source and target speeches in Italian, English and Spanish, named EPIC (European Parliament Interpreting Corpus), is currently being compiled and analyzed (Monti et al. forthcoming; Sandrelli and Bendazzoli 2005 forthcoming).

Corpus-based research is already a well-established branch of Translation Studies, whereas corpus-based interpreting studies as a discipline is still in its infancy. Indeed, several so-called "corpus-based" interpreting studies of the last few years contain analyses of very small sets of data, e.g. one individual interpreter's performance in a single conference (or even part of a conference). Moreover, these studies are based on manual analysis of the data, in that transcripts are not machine-readable: this means that it is not possible to fully exploit the potential offered by the software applications developed by corpus linguists and already widely-used in corpus-based translation studies.

In an influential article published in 1998, Shlesinger (1998a) called for the development of parallel and comparable corpora for interpreting studies. In her opinion, comparable corpora of original and interpreted speeches in the same language could be queried to investigate the characteristics of interpreted language. On the other hand, parallel corpora of source and target speeches could be used to test and validate existing theories on interpreting, particularly as regards interpreters' strategies and interpreting norms. However, researchers interested in creating interpreting corpora are faced with several methodological and practical challenges (Cencini 2002), as is briefly discussed in (2). Section 3 outlines how the EPIC multimedia archive was created, whereas section 4 describes the characteristics of the EPIC corpus. The potential research applications of the EPIC corpus are presented in (5.1), whereas (5.2) explores the teaching applications of the EPIC multimedia archive. Finally, section 6 presents our conclusions and future prospects for the project.

2 Methodological issues in interpreting research

The available literature on interpreting can be broadly categorized either as observational studies or controlled experiments (Gile 1994, 2000; Shlesinger 1998b). Experimental studies are usually carried out to study specific, isolated aspects of interpreting. Clearly, in order for a study to be methodologically sound, experimental conditions and the choice of subjects, materials and tasks are extremely important. It is equally essential that the research goals be clearly defined and delimited from the start.

On the other hand, observational studies offer the chance of seeing all the factors at play in a real working situation, but they are hampered by the problem of collecting sufficient high-quality data, i.e. recordings of conference speeches and interpreted target speeches. Researchers are often refused permission to record conferences by speakers and/or organizers (for confidentiality reasons) and by interpreters themselves, who are not always keen to collaborate in what they perceive as attempts to evaluate the quality of their work (Cencini 2002; Gile 1997; Kalina 1994). However, in the present paper our focus is precisely on collecting and analyzing genuine empirical data from working contexts.

Difficult access to recordings is accompanied by a number of methodological obstacles. The most important ones are the preparation of a rigorous research design and the suitability of the material for analysis, namely its homogeneity.

As regards homogeneity, Shlesinger points out that the complexity of the interpreting process, with its many variables, must be taken into account in order to design a suitable research question and obtain valid results (Shlesinger 1998b:3-4):

> 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.

In a conference interpreting situation the independent variables are manifold. Shlesinger (1998b) presents a review of the available research carried out on the variables related to the modality of interpreting, to speeches, speakers, settings and interpreters. Along similar lines, Alexieva (1997) presents a classification of interpreted events on the basis of broad parameters, including mode of delivery and production, participants, topic, text type and text-building strategies, spatial and temporal constraints and the goal of the event. The present paper does not discuss all of these parameters in detail. It presents merely an overview of the many aspects which must be borne in mind when collecting and analyzing interpreting data.

Some of the independent variables regard the participants in a conference interpreting situation. Gile (1995) talks about a client (the person or organization that requests the professional services of an interpreter), a speaker (or sender), an interpreter, source language listeners and target language listeners (or receivers). Russo (1999) expands the classification suggested by Pöchhacker, who splits the Client role into two different figures, the initiator (the institution or body that organizes the conference) and the client (the party entrusted with the actual organization work), by adding another role, the sponsor, who is seldom physically present in the conference but provides the necessary financial means for the event. All of these figures determine the characteristics of the conference and therefore have an impact on the interpreting service.

The degree of variability is very high: just to give an example, speakers may express their own ideas or be there on behalf of someone else. They may be experienced speakers or first-timers; they may be expressing themselves in their own language or in a foreign language. Their accent may be marked (foreign or regional) or standard. Their style may be formal or casual. They may deliver an impromptu speech or may read a written script, at a fast or slow speed, and so on.

As regards interpreters, their levels of expertise may vary according to training and experience. The quality of the interpreting service may be influenced by their health conditions, preparation (which in turn is influenced by the time and materials available to them before the event), working conditions (the quality of the equipment, noise levels, room layout and visibility conditions from the booths, etc.), and many more factors.

Audience composition is also important, that is, whether the interpreter is translating for a small group of people who are all native speakers of the target language, or whether the service is being provided for a large audience for whom the target language is a foreign language.

Other important variables when analyzing an interpreted event are related to the speech, including the type of event, which in turn influences text types and topics, the degree of technicality of the speech, its information density, input rate (source language speaker's delivery pace), duration, and so on. A very detailed classification of text types comes once again from Alexieva (1994) who identifies several parameters, including mode of production, functional content, use of non-verbal information (visual information, videos, slides, prosodic features, etc.), and degree of intertextuality with the other speeches in the same conference and with external sources.[4]

[4] Another interesting text type classification is the one suggested by Hönig (2002), which was developed for teaching purposes, i.e. to provide a set of objective grades of difficulty of speeches used in training. The model incorporates criteria related to speech topic and structure and students' expected knowledge of the topic in question, as well as cohesion and coherence criteria, speech presentation parameters and parameters related to numbers and figures.

The above cursory glance at some of the variables involved in an interpreted event have highlighted the main obstacles hampering interpreting research. Our research project, which has resulted in the creation of the EPIC corpus and the EPIC multimedia archive, is an attempt to tackle these difficulties by selecting appropriate materials and processing them by using corpus analysis techniques.

The practical problem of access to interpreting data was solved by choosing the European Parliament (EP) plenary sessions. As is explained in more detail in (3.1), the Europe by Satellite (EbS) TV channel enables viewers to select different sound channels for different EU languages. This means that we were able to record both the original speeches and the interpreters' target speeches in Italian, English and Spanish. The material is in the public domain, and EbS authorizes viewers to use it for research and educational purposes. Moreover, the EP debates are published in the verbatim reports available on the European Parliament website, together with information on speakers and speeches (see 3.3).

As regards the methodological aspects mentioned above, the homogeneity of the material under analysis is ensured by the institutionalized setting in which the debates take place. Specific procedures are consistently followed, such as the rules for the allocation of speaking time to MEPs (Members of the European Parliament), the fixed structure of debates, etc. Consequently, text types, topics, speech duration and so on can be controlled when analyzing the material. EP interpreters, for their part, are all professionals who have passed a strict selection procedure, with similar degrees of expertise, if not experience. Moreover, they all work into their "A" language (i.e. into their mother tongue) and they have access to the same information sources to prepare before plenary sessions. Clearly, all EP interpreters enjoy similar working conditions, in that they use the same equipment and booths.

In short, this type of material seemed to offer a high degree of homogeneity and was therefore considered suitable for our research purposes (see 5.1). Section 3 describes how the material was collected and organized in the EPIC multimedia archive.

3 The EPIC multimedia archive

As was explained in section 2, in order to create EPIC it was necessary to collect the materials and organize them in a multimedia archive. The EPIC multimedia archive comprises digital video and audio clips of both original and interpreted speeches. The recorded material was subsequently transcribed for later analysis (see 3.3). The multimedia archive is currently stored on the hard disk of a dedicated machine, but there are plans to upload it to an Internet server to enable external researchers to access the audio and video clips, as well as the EPIC corpus.

Creating the multimedia archive involved a series of challenges, ranging from recording the material, to choosing file formats, storing the files and so on. Moreover, efforts were made in order to keep any extra material that was "unintentionally" recorded but might prove useful for different applications other than research (see 5.2).

3.1 Recording the EP debates

The EP plenary sittings were recorded off the EbS news channel. When the simultaneous interpreting service is available, this satellite TV channel enables viewers to select different sound outputs for different EU languages. Thus, it is possible to listen to the original speakers or to the interpreters working in the various booths.

In our study, four satellite TV + video recorder workstations were used for each plenary sitting to obtain a recording of the original sound channel (abbreviated as "org"), and

recordings of the English, Italian and Spanish sound channels (that is, the interpreters working in the three booths – indicated as "int"). Following the EbS schedule, 4-hour videotapes were used. However, recording sessions were not monitored, resulting in most videotapes containing materials other than EP sittings too. Indeed, as well as the EP part-sessions, the EbS broadcasting schedule includes press conferences and stock footage which European TV channels can use when reporting on EU affairs. In some cases, the simultaneous interpreting service is provided for all EU languages, or it may be restricted to English, French and/or German, or not be available at all. As a consequence, several press conferences were "unintentionally" recorded too. Though such material is not directly linked to the EPIC project, it was decided to store it in a separate archive for pedagogical purposes, as is being done in other institutions (see for example de Manuel Jerez 2003; Carabelli 2003; Gran et al. 2002).

About 2 video tapes per day per language were needed, amounting to 28 tapes for each plenary part-session: overall, 140 videotapes in total, covering 5 part-sessions held from February to July 2004. All the recordings thus obtained had to be digitized, in order to easily select the sections to be studied. The digitization process is still ongoing.

3.2 Digitizing the recordings

The videotapes with the recordings of the original speakers are being digitized as video files, as visual information is potentially useful for later analysis of the corpus. By contrast, the interpreted speeches are digitized as audio files, since the images on the videotapes are exactly the same (i.e. the plenary speakers), whereas our interest lies in audio information (i.e. the interpreters' performances). For each plenary sitting, one video file (the original debate in which all the EU languages may be used as official languages) is thus obtained, together with three audio files containing the same speeches simultaneously interpreted into English, Italian and Spanish.

The original speakers' recordings are converted into digital video files by using *Pinnacle Studio (9.0),* a video-capture and editing software program. The chosen format for the video files is ".mpeg1". The recordings of the interpreted speeches are digitized by using *Cool Edit-Pro 2.0*, a sound editor. The chosen format is the ".wav" format (sample rate = 32.000; channel = mono; resolution = 8 bit), which ensures very good audio quality for possible future studies of prosodic features (distribution of pauses, hesitations, etc.). As was mentioned in (3), there are plans to upload the EPIC archive to a dedicated Web server from which researchers will be able to download the clips. When the project reaches that stage, the ".wav" clips will be converted into a lighter format, probably ".mp3", which would not affect audio quality for users.

When the digitization process of a video tape with the full recording of a plenary sitting has been completed, all the original speeches made in Italian, English and Spanish and their corresponding interpreted versions are selected and saved as individual clips.

Overall, the EPIC multimedia archive includes video clips of all the source language speakers in Italian, English and Spanish, audio clips of the corresponding interpreted versions into two of the three languages involved and the transcripts of both types of spoken material. The transcription process was one of the most challenging parts of the study. Section 3.3 outlines how the material under study was transcribed and provides some useful suggestions on how to ease and speed up the transcription process by using speech recognition programs.

The multimedia archive also contains the full recordings of the part-sessions with speeches potentially in all the other EU languages (depending on the MEPs who took the floor on that particular day), as well as the recordings of a number of press conferences. Some

of the latter feature one or more interpreted versions: all this material is archived and classified for future studies and pedagogical applications.

3.3 Transcription

Once the video and audio digital clips are ready, the material must be transcribed, in order to process and analyze it. As was mentioned in section 1, transcribing spoken material is a demanding and time-consuming task.[5] The material in EPIC is transcribed with a view to building a large amount of data which can then be analyzed automatically, i.e. the transcripts must be machine-readable. However, efforts were made to ensure that the transcripts were also user-friendly, so that anybody would be able to use the EPIC material for their studies. Therefore, we decided to produce very basic orthographic transcripts, with a minimum amount of linguistic and paralinguistic information (Shlesinger 1998a).

We have developed a "transcription procedure" that consists in producing a draft transcript very rapidly, which is then revised several times until it becomes a final draft. In order to produce the preliminary draft of the source speeches, the official verbatim report of each EP sitting, available on the Internet, is used as a basis. The speech features which EP officials routinely correct in the verbatim reports (unfinished sentences, mispronounced words and ungrammatical structures, for example) are re-inserted once again whilst listening to the recordings. Moreover, punctuation is eliminated from the transcripts.

All the target (interpreted) speeches have to be transcribed from scratch. Speech recognition software programs (*Dragon Naturally Speaking* and *IBM Via Voice*) are used to obtain the preliminary drafts which later undergo a revision process. As we are trained conference interpreters, we listen to the recording and repeat aloud what the interpreter says at the same time, that is to say, we apply the shadowing technique (Schweda Nicholson 1990; Lambert 1992). The speech recognition programs are trained to recognize our voices, and produce a draft transcript automatically.

Extra-linguistic data are recorded in a specially-designed header with information about the speech (e.g. duration, mode of delivery, average speed, etc.) and the speaker (e.g. name, nationality, gender, political function, etc) in each transcript. The header fields are also used to set the search parameters in the EPIC web interface (see 4). These search parameters allow users to query only a section of the corpus by selecting speeches on the basis of speech and/or speaker characteristics. This mechanism also allows for fast selection of the materials for teaching purposes (see 5.2).

4 The EPIC corpus

One of the first responses to the practical and methodological challenges described in §2 is EPIC, an interpreting corpus in the technical sense of the word (as used in corpus linguistics), that is to say, an electronic collection of transcripts of European Parliament speeches and their interpreted versions, in three languages (Italian, English and Spanish). The speeches were delivered during EP plenary sittings by MEPs, by the EP President and Vice Presidents, by European Commission and European Council representatives, and by guests from EU and non-EU countries. The interpreted speeches were produced by the EP interpreters working during those sittings.

[5] A fuller account of transcription methods and conventions used in EPIC can be found in Monti et al. (forthcoming).

EPIC is an open corpus, in that it will be expanding over time as more data is added to it. By date, one part-session (February 2004) is available for study, corresponding to about 18 hours of transcribed material. Other part-sessions (two in March, one in April and one in July 2004) are being processed and will be added to the corpus as they become available.

As was previously mentioned, EPIC is a trilingual corpus and each source speech in one language (from among Italian, English and Spanish) is accompanied by the corresponding target versions in the other two languages. In this sense, EPIC is not a single corpus, but is made up of a collection of 9 sub-corpora, namely 3 sub-corpora of source texts (original speeches) and 6 sub-corpora of target texts (simultaneously interpreted speeches), to which 6 sub-corpora of aligned texts will be added as a next step in the project. Table 1 shows the structure of the corpus and its present size:

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%**

Tab. 1 Composition of EPIC

Finally, and most importantly, EPIC is machine-readable. In order to make automatic analysis possible, the corpus is POS-tagged and lemmatized by using specific taggers, namely the *Treetagger* for the English language, a combination of taggers proposed by Baroni et al. (2004) for the Italian language and *Freeling* for the Spanish material. Then, the material is encoded by using the *IMS Corpus Work Bench* platform (Christ 1994), which enables users to carry out simple and advanced queries by using the *CQP* query language of *CWB* (Bendazzoli et al 2004; Monti et al. forthcoming). The corpus can be queried through a dedicated web interface that is available on the Forlì School for Translators and Interpreters' development website, hosting a number of resources for linguists, translators and terminologists[6]:

The EPIC web interface enables users to carry out queries to retrieve and analyze material of interest, either in the whole corpus or by restricting the search through the use of speech- and speaker-related search filters. Thanks to these filters, based on the header fields contained in each transcript, it is possible to restrict queries on the basis of specific characteristics, such as speakers' political function, country of origin, speech mode of delivery, speech length, duration, and so on.[7]

[6] http://sslmitdev-online.sslmit.unibo.it/corpora/corpora.php

[7] For more details on the tagging process and the development of the EPIC web site, as well as the available search filters, see Monti et al. (forthcoming).

5 Research and pedagogical applications

5.1 Research: corpus-based interpreting studies

Electronic corpora have long been awaited in Interpreting Studies in order to validate the many hypotheses and theories suggested by scholars on interpreters' strategies and the interpreting process (Shlesinger 1998a). Most interpreting research is still based on small case studies which are conducted through manual analysis or exploit semi-automatic analysis in very limited terms. Moreover, the creation of a multilingual parallel corpus of interpreted speeches and their corresponding source speeches also offers the opportunity of comparing more translations of the same text, something which is not often possible, as pointed out by Kalina (1994:227): "In studying real-life conditions and professional interpeting [sic], one problem is that one will rarely find several interpreted versions of the same text, a fact which makes direct comparison impossible".

Once EPIC reaches larger dimensions and the various sub-corpora are of matching sizes, it will be possible to provide results that are based on statistical measures and corroborate hypotheses on the basis of a significant number of occurrences of a given feature. Moreover, since EPIC transcripts are tagged, lemmatized and encoded, they can be searched not only on the basis of word forms, but also on the basis of the corresponding parts of speech, lemmas and possible pronunciation disfluencies.

The present level of transcription and annotation offers various research opportunities, such as studying lexical patterns, frequency lists, concordances (Partington 2001:47), collocations , use of prefixes and suffixes, and so on. Generally speaking, various lines of research already developed in corpus-based translation studies (see for example Bowker 2002; Laviosa 2002) could be followed in exploring the corpus. Incidentally, such explorations of the EPIC corpus may not only reveal interesting characteristics of the material itself, but may also shed light on the differences between translation and interpreting.

EPIC can be explored either as a parallel or a comparable corpus. As regards EPIC as a parallel corpus, the next step in the project envisages the content alignment of source and target speeches. Thus, six more sub-corpora of aligned speeches will become available, making it possible to carry out semi-automatic queries using the web interface already described above. The aligned corpora will make it easier to carry out studies on quality features and specific interpreting strategies. Moreover, considering the main research interest of the group, i.e. directionality, the multilingual and multidirectional nature of the EPIC corpus will enable us to focus on the possible differences depending on the language pair (between two Romance languages or one Romance and one Germanic language) and language direction (from a foreign language into the native language or vice versa). As highlighted by Johansson (1998:6-7), "to distinguish between what is language-specific, and what is general, it is useful to turn to translations of the same source texts into a variety of languages".

If EPIC is explored as a comparable corpus, the 9 sub-corpora of original and interpreted speeches can be grouped on the basis of language, thus allowing for studies that compare original English, for example, and the English used by simultaneous interpreters. "Interpreted English" can then be further analyzed to see whether there are any differences depending on the source language (in our case, either Italian or Spanish). Indeed, the first attempt to explore EPIC and exploit its research potential using semi-automatic analysis was a study on lexical patterns in simultaneous interpreting (Sandrelli and Bendazzoli 2005 forthcoming) to verify Laviosa's findings on non-translational (original) and translational English (Laviosa 1998) and see whether her results also apply to original and interpreted

English and original and interpreted Italian.[8] As this was the very first exploration, great efforts had to be made to correct unexpected flaws in the system and master the necessary techniques of analysis.

Finally, it must be pointed out that, given the user-friendly nature of our transcripts, it should be fairly easy to add further levels of annotation, such as linguistic, paralinguistic or extra-linguistic features (Leech 1997:5). Examples include pauses, false starts, syntax, prosodic features and even speakers' body language, humor, etc.

5.2 Pedagogical applications: multimedia archive materials

The potential pedagogical applications of the EPIC multimedia archive concern both foreign language teaching (especially for L2 students) and interpreter training. Source speeches are potentially of interest to both groups of users, whereas interpreted speeches are a useful resource for trainee interpreters.

As regards foreign language teaching, the video clips of the source speeches can be used for listening comprehension exercises. The availability of the speech transcripts can further help students, who can read the text after listening to the speech and then focus their attention on any unknown words or structures.

Moreover, listening exercises are also useful for improving students' pronunciation skills in the foreign language. An example of pronunciation exercises based on the archive (which, as was explained in (3), includes recordings in all the EU languages) is the series of German materials for the SPT-Sound Perception Trainer course aimed at the students of the Forlì School for Interpreters and Translators (Kaunzner 1997).

A potential application which may be of interest to both L2 learners and trainee interpreters is the comparative study of rhetorical devices employed in EP speeches. The availability of source speeches in the three languages offers students the opportunity to compare the different rhetorical devices and stock phrases used in English, Italian and Spanish formal speeches, and, more specifically, the special features of language typical of the EP context (EU jargon and other conventions). Moreover, the verbatim reports published on the EP website can be compared with the actual transcripts of the source speeches, in order to identify the main differences between spoken texts and polished written texts.

In this regard, the EPIC corpus, as well as the multimedia archive, can be used as a teaching tool, in that it enables students to carry out targeted searches for specific structures and expressions in all three languages, along the lines suggested by Zorzi (2001), who discusses how to teach the use of discourse markers by using spoken corpora. Moreover, the "topic" search parameter enables students to study the features of speeches by topic, that is, on politics, economics, health, and so on. The "Procedure & Formalities" option may be of particular relevance to trainee interpreters who can use it to study the specific formulaic language used in all of their working languages in the EP context.

Two applications which are more directly relevant to interpreter training are the use of EPIC video clips and transcripts as practice materials, and the use of the EP interpreters' target speeches for self-assessment purposes. As regards the former, EPIC source speeches may be used by trainers during classes to present students with real-life assignments. The speech classification system implemented in the headers of the transcripts and searchable via the EPIC Web interface is a useful source of information for teachers when selecting class materials. In particular, teachers may choose speeches by speed, topic, accent, etc. If a selected clip is considered too difficult to interpret for the specific stage reached by the

[8] The Spanish materials in EPIC will be included in a future study.

students, it can be edited by using *Cool Edit* or similar software tools, for example to divide it into several clips, to slow it down without altering the speaker's pitch, to insert pauses in the speech, etc. Moreover, the type of materials available through EPIC are ideal for use in any CAIT (Computer Assisted Interpreter Training) software tool (cf. Sandrelli 2007, Sandrelli 2003a, 2003b; Carabelli 1999, 2003; Gran Tarabocchia et al. 2002).

The target speeches produced by EP interpreters may find an application in the training of student interpreters too, in that they offer a useful demonstration of professional interpreting standards. Students may be asked to interpret a speech from the archive, either in class or in their individual study time. The recording of their performance can then be compared with the corresponding professional interpretation available in the archive. The assessment exercise may be carried out in class together with other students and under the guidance of a teacher (co-assessment), or in privacy (self-assessment). In both cases, it may contribute to enhancing students' awareness of their strengths and weaknesses, thus giving them useful indications for future work.

6 Conclusions

The Directionality Research Group was originally set up to carry out a research project on directionality in interpreting. In its first year of activity, the group has produced two tools: the EPIC multimedia archive and the EPIC corpus, which is available to the whole research community on a dedicated web page.

As was mentioned in (3) and (4), work is continuing to expand both tools. The multimedia archive, created as a source of materials to be transcribed and analyzed via corpus linguistics techniques, has turned out to hold great multidimensional potential for teaching purposes, as was outlined in (5.2). The archive can also be integrated with other materials for teaching purposes as well, such as pages from the EP website, which provides a wealth of extra information about speakers and debated issues.

The EPIC corpus, on the other hand, is the first publicly available corpus of original and interpreted speeches in three European languages. It is hoped that its exploration will yield interesting results which will contribute to interpreting research on general, language-specific and directionality-related interpreting strategies, and at the same time will inform about teaching methods.

7 References

Alexieva, Bistra (1994): 'Types of Texts and Intertextuality in Simultaneous Interpreting', in M. Snell-Hornby, F. Pöchhacker and K. Kaindl (eds) *Translation Studies. An Interdiscipline*, Amsterdam & Philadelphia: John Benjamins, 179-187.

------ (1997): 'A Typology of Interpreter-Mediated Events', *The translator* 3(2): 153-174.

Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston and Marco Mazzoleni (2004) 'Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian', in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa 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*, 5: 1771-1774.

Bendazzoli, Claudio, Cristina Monti, Annalisa Sandrelli, Mariachiara Russo, Marco Baroni, Silvia Bernardini, Gabriele Mack, Elio Ballardini and Peter 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, Proceedings of the 4th International Conference on Language Resources and Evaluation*: 33-39.

Bowker, Lynne (2002): *Computer-Aided Translation Technology: a Practical Introduction*, Ottawa: University of Ottawa Press.

Carabelli, Angela (1999): 'Multimedia Technologies for the Use of Interpreters and Translators', *The Interpreters' Newsletter* 9: 149-155.

------ (2003) 'A Brief Overview of IRIS – The Interpreter's Research Information System', in J. Jerez de Manuel (Coord) *Nuevas Tecnologías y Formación de Intérpretes*, Granada: Editorial Atrio, 113-139.

Cencini, Marco (2002): 'On the Importance of an Encoding Standard for Corpus-Based Interpreting Studies. Extending the TEI Scheme', *inTRAlinea Special Issue: CULT2K*, available online at 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 2004*.

de Manuel Jerez, J. (2003): 'Nuevas Tecnologías y Selección de Contenidos: la Base de Datos *Marius*', in J. Jerez de Manuel (Coord) *Nuevas Tecnologías y Formación de Intérpretes*, Granada: Editorial Atrio, 21-61.

Gile, Daniel (1994): 'Methodological Aspects of Interpretation and Translation Research', in S. Lambert and B. Moser-Mercer (eds) *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, Amsterdam & Philadelphia: John Benjamins, 39-56.

------ (1995): *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam & Philadelphia: John Benjamins.

------ (1997): 'Interpretation Research: Realistic Expectations', in K. Klaudy 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, 43-51.

------ (2000): 'Issues in Interdisciplinary Research into Conference Interpreting', in B. Englund Dimitrova and K. Hyltenstam (eds) *Language Processing and Simultaneous Interpreting: Interdisciplinary Perspectives*, Amsterdam & Philadelphia: John Benjamins, 89-106.

Gran Tarabocchia, Laura, Angela Carabelli and Raffaela Merlini (2002): 'Computer-Assisted Interpreter Training', in G. Garzone and M. Viezzi (eds) *Interpreting in the 21st Century. Challenges and Opportunities*, Amsterdam & Philadelphia: John Benjamins.

Hönig, Hans G. (2002): 'Piece of Cake - or Hard to Take? Objective Grades of Difficulty of Speeches Used in Interpreting Training', in *Teaching Simultaneous Interpretation into a "B" Language, EMCI Workshop 20-21 September 2002*.

Johansson, Stig (1998): 'On the Role of Corpora in Cross-Linguistic Research', in S. Johansson and S. Oksefjell (eds) *Corpora and Cross-Linguistic Research. Theory, Method and Case Studies*, Amsterdam & Atlanta: Rodopi, 3-24.

Kalina, Sylvia (1994): 'Analyzing Interpreters' Performance: Methods and Problems', in C. Dollerup and A. Loddegaard (eds) *Teaching Translation and Interpreting 2: Insights, Aims, Visions*, Amsterdam & Philadelphia: John Benjamins, 225-232.

Kaunzner, Ulrike A. (1997): 'Audio-Lingua: Pronunciation Improvement Through Sound Perception Training', available on-line at
http://www.tomatis.se/tomatis/tomatis.nsf/8db568a96c37cd9e8625674e00714a92/8cc4b812253cec6cc125687e0035ce84/$FILE/Language%20Training.htm (22/07/2005)

Lambert, Sylvie (1992) 'Shadowing', *The Interpreters' Newsletter* 4: 15-24.

Laviosa, Sara (1998): 'Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose', *Meta* 43(4): 557-570.

------ (2002): *Corpus-based Translation Studies: Theory, Findings, Applications*, Amsterdam & New York: Rodopi.

Leech, Geoffrey (1997): 'Introducing Corpus Annotation', in R. Garside, G. Leech and T. Mc Enery (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Longman, 1-18.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli and Mariachiara Russo (forthcoming) 'Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus)', *Meta*.

Partington, Alan (2001): 'Corpora and their Uses in Language Research', in G. Aston (ed.) *Learning with Corpora*, Bologna: Clueb, 46-62.

Russo, Mariachiara (1999): 'La Conferenza come Evento Comunicativo', in C. Falbo, M. Russo and F. Straniero Sergio (a cura di) *Interpretazione Simultanea e Consecutiva. Problemi Teorici e Metodologie Didattiche*, Milano: Editore Ulrico Hoepli, 89-102.

Sandrelli, Annalisa (2003a): 'Herramientas Informáticas para la Formación de Intérpretes: Interpretations y The Black Box', in J. de Manuel Jerez (Coord) *Nuevas Tecnologías y Formación de Intérpretes*, Granada: Editorial Atrio, 67- 112.

------ (2003b): 'New Technologies in Interpreter Training: CAIT', in H. Gerzymisch-Arbogast, E. Hajičová & P. Sgall, Z. Jettmarová, A. Rothkegel and D. Rothfuß-Bastian (eds), *Textologie und Translation, Jahrbuch Übersetzen und Dolmetschen 4/II*, Tübingen: Gunter Narr Verlag, 261-293.

------ (2007) 'Designing CAIT (Computer-Assisted Interpreter Training) tools: *Black Box*'. Proceedings of the Marie Curie Euroconferences MuTra '*Challenges of Multidimensional Translation'* - Saarbrücken 2-6 May 2005.

------ and Claudio Bendazzoli (2005 forthcoming): 'Lexical Patterns in Simultaneous Interpreting a Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)', in *Proceedings from the Corpus Linguistics Conference Series*, 1(1), ISSN 1747-9398, forthcoming on-line at www.corpus.bham.ac.uk/PCLC

Schweda Nicholson, Nancy (1990): 'The Role of Shadowing in Interpreter Training', *The Interpreters' Newsletter* 3: 33-40.

Shlesinger, Miriam (1998a): 'Corpus-Based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies', *Meta* 43(4): 486-493.

------ (1998b): 'Interpreting as a Cognitive Process: What Do We Know About How It Is Done?', *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.

Zorzi, Daniela (2001): 'The Pedagogic Use of Spoken Corpora: Learning Discourse Markers in Italian', in G. Aston (ed.) *Learning with Corpora*, Bologna: Clueb, 85-107.

Web references

EbS (Europe by Satellite): http://www.europa.eu.int/comm/dg10/ebs
EPIC interface: 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 Work Bench, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
TreeTagger:
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html