INTRODUCING CORPORA IN INTERPRETING STUDIES: FROM EPIC TO DIRSI

Claudio Bendazzoli
University of Turin (Italy)
School of Management and Economics
Department of Economic and Social Studies

Abstract: In this paper I briefly illustrate the development of the corpus-based approach to the study of spoken language interpreting. Drawing on the seminal works by Miriam Shlesinger (1998) and Robin Setton (2011), I describe the main challenges and opportunities that have been found in interpreting corpora, from the early studies, based on small data sets analyzed manually, to more recent ones in which larger, machine-readable corpora have been created. In particular, I present two major corpus-based projects concerning simultaneous interpreting at the European Parliament and at international medical conferences held in Italy. The two projects led to the creation of the European Parliament Interpreting Corpus (EPIC) and the Directionality in Simultaneous Interpreting Corpus (DIRSI) respectively. Similar methodological challenges can be related to each setting, but the fundamental differences in data access and the specific features of the communicative situations considered provide for alternative solutions. These are described together with the opportunities afforded by interpreting corpora in terms of research, teaching, and professional practice.

Key Words: corpus-based research; conference interpreting; simultaneous interpreting; methodology.

1. The corpus-based approach

Computer technology has been having a profound impact on the development of different academic fields, including linguistics. For linguists, the computer and the Internet have paved the way to analyzing larger datasets in a systematic fashion, to the extent that it would be impossible to process them with the naked eye (Biber et al. 1998). As reported by Laviosa (2011), such an unprecedented opportunity found fertile ground in (written) translation studies at the beginning of the 1990s (e.g. Baker, 1993), became an established research paradigm in the second half of the same decade, and then started to spread across languages and cultures from the beginning of the new millennium. This expansion continues to open new research directions (Fantinuoli and Zanettin, 2015) and has also begun to inform translation training and practice (Bernardini and Castagnoli, 2008).

Due to the intrinsic difficulties in gathering, transcribing and making spoken (and sign-language) data available in electronic form, the corpus-based approach began to be considered some time later in interpreting studies. Regrettably, this gap between corpus-based translation research and corpus-based interpreting research has yet to be closed, as is evident in the still limited examples of corpus use in interpreter training, not to mention interpreting practice. Nevertheless, much progress has been made since interpreting scholars started considering this approach, as can be appreciated in the works by Shlesinger (1998) and Setton (2011) which can be considered milestones in the development of this paradigm.

CIS: The Beginning

Probably the first paper about the idea of extending the corpus-based methodology to interpreting, as well as using already available monolingual corpora in experimental studies was published by Miriam Shlesinger in 1998. In this seminal work, Shlesinger emphasizes "the potential to use large, machine-readable corpora to arrive at global inferences about the interpreted text" (*ibid.*: 2), taking inspiration from years of successful research in corpus-based translation studies.

Transcribing spoken language data and representing relevant nonverbal features are singled out as critical obstacles to the creation (and distribution) of large corpora. Notwithstanding these drawbacks, which have been tackled in a number of ways (see below), Shlesinger promotes the use of interpreting corpora as they can be valuable tools within the realm of descriptive studies. In particular, multiple research opportunities may come from the use of comparable or parallel corpora. The former compare different data sets produced in the same language but in different conditions (e.g. as original speeches; as interpreted speeches using different modes;

as written source texts; as translations); the latter compare source texts or speeches and their target versions as a result of the interpreting (or translation) process. The two perspectives can yield insight into general and specific features of language production, translation modes and language direction. In addition, experimental research may also benefit from the use of corpora. As suggested by Shlesinger, already available corpora can be queried to obtain source texts (or particular strings of texts) for running experiments and testing hypotheses.

A Flourishing Offshoot

More than ten years after Shlesinger's paper, in which CIS are seen "as an offshoot of corpus-based translation studies" (*ibid.*), Robin Setton (2011) published a chapter which provides a broad overview of several corpus-based research projects, including more than 20 works across several language combinations, directions, and modes. Some of these projects were carried out long before Shlesinger's call – examples of this kind range from Oléron and Nanpon (1965) to Pöchhacker (1994) and in some cases there is no more availability of transcripts or recordings. In fact, part of the projects reported in the overview are based on manual analysis and do not take advantage of automatic extraction of occurrences as is normally the case in corpus linguistics-aided investigations. Overall, Setton's overview shows that the notion of corpus[1] was initially upheld to underscore the empirical nature of these studies, which are based on authentic instances of interpreter-mediated communication and not on introspection alone or artificially generated stimuli. As outlined in Bendazzoli and Sandrelli (2009), the development of CIS includes such initial efforts still oriented to manual analysis, early machine-readable corpora (with digital recordings and transcripts in electronic form), and fully-fledged electronic corpora (indexed for computerized queries). The two CIS projects presented below in greater detail belong to the third group of corpora. Both EPIC and DIRSI are available in electronic form and have been designed in a way to be accessible (or easily adjustable) via multiple query tools.

2. From EPIC to DIRSI: similar challenges, different solutions

In this section I compare and contrast the methodological choices made to overcome similar obstacles in the creation of two interpreting corpora, i.e. EPIC and DIRSI. The obstacles under consideration are related to the following steps in corpus building:

1. Corpus design
2. Data collection
3. Transcription
4. Markup and annotation
5. Alignment
6. Access

These steps are adapted from the guidelines put forward by Thompson (2005) with reference to the creation of spoken language corpora. Thompson's guidelines only include steps 2, 3, 4, and 6, and were supplemented with further steps on the basis of the experience gained in the creation of the two corpora mentioned above.

It must be specified that the two projects started at different times and on different assumptions. The EPIC project began in 2004 and involved a research group of scholars with expertise in a number of fields, such as interpreting studies, translation studies, and computational linguistics, with two to three members dedicated full time to it (Monti et al. 2005; Russo et al. 2012). The DIRSI corpus is the result of a PhD project which lasted four years (Bendazzoli, 2010; 2012) with the collaboration of many colleagues, in particular the Computational Linguistics Lab of the Universidad Autónoma in Madrid (Spain). In both projects the focus is on professional simultaneous interpreting between spoken languages.

Corpus design

Both EPIC and DIRSI are in fact a collection of multiple sub-corpora representing original and interpreted language. EPIC is a trilingual corpus and consists of 9 sub-corpora: 3 sub-corpora of source speeches (Italian, English, Spanish) and 6 sub-corpora of target speeches as a result of all the possible combinations and directions between the three languages involved; DIRSI is a bilingual corpus and includes 4 sub-corpora: 2 sub-corpora of source speeches in Italian and English and 2 sub-corpora of relevant target speeches into English and into Italian.

It is clear that the number of languages, together with the Translation modality and modes represented in a corpus impact on the overall design and structure of the corpus itself. Typically, when simultaneous interpreting is provided, participants are required to use a microphone, speak one at a time, and follow a commonly shared protocol. Failing to comply with these "rules" would create disruption in the interpreter-mediated communication and is likely to force the interpreters to warn the audience and the speakers about this. Given the monologic and, above all, institutionalized (Heritage, 1995: 408) nature of the speech events recorded in EPIC and DIRSI, each participant's turn would match the beginning and the end of a ratified speech event, which could be treated separately from the speech events preceding and following it in constructing the corpus. On the other hand, in case of really dialogic interaction (in the sense of spontaneous and not as pre-organized beforehand as is the case in parliamentary debates and conference proceedings), a similar separation of each participant's turn is hardly manageable and a larger section of the communicative event must be considered to account for possible overlapping, latching, and other conversational features typical of unregulated (or simply less regulated) talk where the floor is "locally managed" (Sacks et al. 1974: 725). This is the reason why Q&A sessions are not represented in DIRSI. Though they form part of the conference proceedings under consideration, they had to be excluded due to the highly interactive and quasi-spontaneous nature of many of the exchanges that took place in them. Conversely, EP debates are much more regulated with precise rules for the allocation of speaking time [2] and management of the floor as "single person floor" and "one prime speaker floor" (Hayashi, 1996: 70-71). There are cases of overlapping speech, but these are more the exception than the rule.

The inclusion of audio/video recordings for possible alignment to the textual part of the corpus also plays a significant role in shaping the organization of materials to be subsequently queried or referenced to. It is thus necessary to adopt a coherent system in naming the different files, so as to retrieve them easily and to be able to gain some basic information just by reading the file names. The following examples show the naming system adopted in EPIC and DIRSI respectively. Table 1 and Table 2 detail the items included in each system. Example 1 is taken from an EPIC source text delivered in English on 10 February 2014, in the morning sitting, which is interpreted into Italian and into Spanish, and can be found at the beginning of the official verbatim report for that day (since it is assigned number 005).

Example 1:
10-02-04-m-005-org-en
10-02-04-m-005-int-en-it
10-02-04-m-005-int-en-es

Table 1: File Naming System in EPIC

Date	DD-MM-YY
Morning / afternoon sitting	m / p
Reference number from verbatim report[3]	000
Text type **(original or interpretation)**	org / int
Language **(for source texts only)**	it / en / es
Language direction **(for target texts only)**	int-en-it / int-it-en / int-es-it int-en-es / int-it-es / int-es-en

The system used to name DIRSI files is largely based on the one used for EPIC, but with some adjustments and further additions. In Example 2, the corpus name is also displayed; the date is written in reverse order; city and conference codes are added to identify quickly the materials in the multimedia archive; the language direction also provides information about the directionality: small caps are used to indicate the interpreter's B language (i.e. active working language) whereas capital letters are used to indicate the interpreter's A language (i.e. native language). Since all the interpreters represented in DIRSI work in both directions[4], this distinguishing feature needed to be recorded and displayed.

Example 2:
DIRSI-2006-05-20-VR-CFF4-001-org-it
DIRSI-2006-05-20-VR-CFF4-001-int-it-EN

Table 2: File Naming System in DIRSI

Corpus name	DIRSI
Date	YYYY-MM-DD
City code	(VR, FC, etc.)
Conference code	(CFF4, CFF5, ELSA, etc.)
Progressive number[5]	000
Text type **(original or interpretation)**	org / int
Language **(for source texts only)**	it / en
Language direction and directionality **(for target texts only)**	int-en-IT / int-IT-en int EN-it / int-it-EN

Despite being the first step in the corpus compilation process, many of the methodological choices made to design the two corpora could be fully defined only after transcribing some of the recordings and becoming more familiar with the data, thus revisiting previous steps already made in the initial part of the projects.

Data collection

Gathering data from interpreter-mediated communicative situations has always been a difficult task. However, technological innovation and better informed approaches have made it easier for scholars to obtain data either directly or indirectly. The world wide web, for instance, now gives access to a wide range of conferences, festivals (e.g. Bani, 2016), lectures, and other events mediated by interpreters. This is the case of the European Parliament and its video library, in which all the plenary sittings held since April 2006 are stored and can be downloaded. Unfortunately, this incredible resource was not available at the time of the EPIC project, so it was necessary to videorecord EP debates broadcast by satellite TV channel EbS. Four TV sets were used (one to record the original channel, one to record the English booth, one for the Italian booth and one for the Spanish booth). In total, 140 videotapes (lasting 4 hours each) were collected after recording the plenary sittings held in February, March, April, and July 2004. All the videotapes were then digitized and the resulting files were edited to extract the clips of the relevant speeches delivered in the three languages under study. Despite having indirect access, plenty of contextual information could be gained about the communicative situation at stake, the participants involved, the procedures regulating the allocation of speaking time and, more generally, the organization of the debates.

In order to collect data from international conferences in Italy, direct access to the communicative situation was deemed the best option. This entailed an ethnographic approach through fieldwork and direct involvement as a practisearcher (Gile ,1994; see also Bendazzoli, 2016). Recording equipment consisted of two laptops, one connected to the conference room audio system (for the source speeches) and one inside the booth (alternatively a small digital recorder could be used). Contextual information had to be gathered on the spot and through the conference organizers, so as to know more about the participants and the content of their presentations.

Being involved in most of these events both as a fieldwork researcher and as a practicing interpreter has advantages and drawbacks. The main advantage is to be granted full access to the contents, the participants and, most importantly, to the backstage of the communicative situation. However, one needs to be well trained to be able to oversee all the research-related tasks and the assignment-related ones. In DIRSI only audiorecordings were obtained, though the size and memory capacity of current cameras are likely to ease even more the collection of video data in most settings.

It is important to specify that the principle of transparency in the European Parliament and the public nature of the events recorded for DIRSI did not pose serious problems of confidentiality. Other settings may raise greater obstacles due to the need to anonymize participants or pieces of information, and this should be reflected in a more detailed consent form which could actually discourage participants from taking part in a study (Metzger and Roy, 2011).

Transcription

For a spoken (or sign language) corpus to be suitable for analysis, be it with or without computerized tools, the data must be transcribed. Of all the necessary activities to compile a corpus, transcribing remains time-consuming and has a bearing on the overall size of interpreting corpora. In fact, large spoken corpora do exist, e.g. the spoken part of the British National Corpus is 10 million words and is now being further expanded and updated (Dembry and Love, 2015). However, very large corpora projects such as the BNC usually count on considerable funding and teams of professional transcribers. This is not the case in interpreting studies, as only small, or relatively small, and specialized corpora have been created so far. It is also true that the extent to which communication is mediated by an interpreter is far more limited compared to non-mediated communication in absolute terms, but this should not discourage interpreting scholars from aiming at bigger data. In this respect, the European Parliament is an impressive source of data, though the results obtained from looking at this setting should not be generalized to represent what happens in simultaneous interpreting also in other settings.

A fundamental observation to bear in mind when transcribing spoken language is that the final product, i.e. the transcript, is a selection of features from the verbal and nonverbal dimensions of communication, and this selection already implies methodological and theoretical choices on the part of the transcriber (Brown and Yule, 1983: 11ff.). Given the great amount of data recorded in the EPIC project, it was clear from the very beginning that it was not possible to represent many nonverbal features. Thus an orthographic transcription with a basic level of annotation was devised (see below).

Moreover, to speed up the transcription process, the verbatim reports of the debates were used as a first draft, which was then revised to fully match the spoken delivery of each source speech; the target texts produced by the interpreters were transcribed by means of speech recognition software, with the transcriber listening to the recording and repeating it aloud (the three transcribers involved were all trained in simultaneous interpreting and were thus able to perform ‘shadowing’). Speech recognition software was also used to transcribe the recordings for DIRSI, but in this case there was no verbatim report to be used as a first draft for the source speeches. The availability of the conference programs, the power point presentations and field notes came particularly handy to disambiguate unclear names or technical terms. As pointed out earlier, even the transcription procedure could be fully defined after dealing with part of the data using a sort of trial and error approach. In EPIC, the availability of the verbatim reports guided some of the methodological choices, which were then carried over to the procedure developed in DIRSI. More specifically, a first draft of all the speech events in a conference session was made first, so as to number each of them and obtain an overall map of the session. This served as a kind of verbatim report in which the types of speech events (e.g. conference presentations, opening and closing remarks, floor management and so on) and the roles of the participants (e.g. chair person, presenter or lecturer, discussant, etc.) could be identified.

Markup and annotation

An annotated corpus is far richer in information than a corpus without any sort of annotation, which "can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process: the linguistic symbols which are attached to, linked with, or interspersed with the electronic representation of the language material itself" (Leech, 1997:2).

Three levels of annotation were considered in EPIC and DIRSI, namely a verbal or linguistic level, a nonverbal (paralinguistic and kinetic) level, and an extra-linguistic or contextual level (metadata). As mentioned earlier, the transcription process is selective, therefore only a limited number of features were included in each annotation level. In particular, the nonverbal level was kept to a minimum with the sole inclusion of empty and filled pauses (this feature was not maintained in DIRSI), mispronounced words and unfinished words. For the extralinguistic level instead a few entries were defined in the form of a header at the beginning of each transcript. The header contains metadata such as details of the file, the speaker, the speech event, and so on. It was not possible to replicate the EPIC header in DIRSI due to the distinguishing features of the two settings, and to further advancement in the theoretical reflection (for more details, see Bendazzoli, 2012). Both corpora were also POS-tagged and lemmatized[6] with Treetagger (Schmid, 1994; 1995). In DIRSI, manual correction of some tags was also performed (e.g. conjunction vs. relative pronoun "that" / "che"). Furthermore, in order to align the transcripts to the audio recordings, time tags were manually inserted by means of transcription software Transana.[7] These tags annotate time information in milliseconds and one can then simply click on a section of the transcript to listen to that particular part of the recording. Finally, both corpora were indexed via the IMS Corpus Work Bench – CWB (Christ, 1994).

A "plain" version of the transcripts was also kept in pure text form to be able to annotate, index and query them with other tools, or simply to use them for pedagogical purposes.

Alignment

Just like annotation, alignment is not really compulsory for an interpreting corpus but, if performed, it enhances the research potential considerably. Alignment can be referred to two different areas. The first area concerns the alignment of the transcripts to the relevant video or audio files, so that one can watch or listen to a fuller representation of the data and supplement what is made available in written form. The second area regards the alignment of source and target speeches, which can be displayed in a parallel fashion on the basis of content or on the basis of *décalage* (for simultaneous interpreting).[8]

This step in corpus development was not made in the EPIC project, though the file naming system provides convenient references to find transcript and multimedia files. Since the duration of most speeches in EPIC is under 10 minutes, it is easy for example to import them into a software program like Transana and have them displayed in three parallel columns.

In DIRSI, both areas of alignment were attained. The transcribed files in this corpus are aligned to the relevant audio files, and each source text transcript is aligned to its target text transcript on the basis of content.

Access

The two corpora are openly accessible to the research community at large, provided that no commercial use is made and that distribution is always referred to the original authors. EPIC has an online interface[9] with simple and advanced query options to extract occurrences from the annotated transcripts only. Multimedia files, as well as the tagged transcripts, can be obtained from the European Language Resources Association catalogue.[10] DIRSI is also available online on a website[11] hosted in a server of the Computational Linguistics Lab of the Universidad Autónoma of Madrid. The website gives access to all the aligned transcripts and to a concordancer.

The format of the transcripts makes it possible to easily adjust them for use with other corpus linguistics tools, thus widening the scope of the research approaches and the angles from which these data are observed and described.

3. CONCLUDING REMARKS

Though slower than corpus-based translation studies, corpus-based interpreting studies have been developing constantly ever since they were first advocated by Miriam Shlesinger at the end of 1990s. In fact, the very notion of ‘corpus’ has been applied in a flexible way to include not only large and machine-readable data sets, but also collections of data that are put together according to inclusion and exclusion criteria to achieve a representative sample but are only suitable for manual analysis. This broad definition of corpus in interpreting research has remained tenable due to the relatively limited size of most interpreting corpora, and will continue to be so insofar as corpus size remains manageable through more traditional approaches. Notwithstanding this fundamental difference, the use of corpora has marked a crucial shift to more empirical and descriptive research.

The corpus development process is also helpful to focus on many critical steps present in empirical research, in that “special challenges” are posed by the “added dimensions of interpreting – multilingualism, orality, situatedness and immediacy” (Setton, 2011: 68). Each step entails a number of specific challenges, such as transcription, annotation, alignment and so on, with multiple solutions. For instance, given the difficulties to achieve a common standard in transcription, it would be important to conform at least to a basic level of annotation to allow for more data sharing and comparability. This is especially relevant to rich data sources, such as the European Parliament or the governmental press conferences broadcast on Chinese television, which are being used by more research teams in different parts of the world (e.g. Wang, 2012a, 2012b; Hu and Tao, 2013). Whatever the granularity of annotation achieved in a corpus, it is important to keep a set of transcripts without annotations (or at least with a minimum level of annotation) to be able to use them with alternative tools. Software programs are capable of processing, counting, and extracting occurrences to an extent that would not be possible through more traditional methods. However, it is probably by combining quantitative analyses with qualitative ones that a fuller picture can emerge from interpreting research. Corpora must be seen for what they are, i.e. sources of information whose representativeness and reliability much depend on the methodological choices made by the researchers who put them together. Moreover, interpreting corpora are invaluable resources to be exploited not only in research, but also in interpreter education (e.g. Russo, 2010; Sandrelli, 2010) and in professional practice. These two areas have yet to be explored extensively under this paradigm and it is hoped that interpreting corpora will be introduced in them soon.

NOTES

[1] In corpus linguistics, a corpus is defined as "a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria" (Bowker and Pearson, 2002: 9).

[2] See, for instance, Rules 162, 163, and 164 of the EP Rules of Procedure (these are available online in all EU official languages: http://www.europarl.europa.eu/sides/getLastRules.do?reference=ANN-09&language=EN).

[3] In all the verbatim reports downloaded from the EP website in February, March, and April 2004, each MEPs intervention is assigned a progressive number which was used as a reference in the corpus as well. As of the July 2004 sitting, this feature was not present in the reports and numbers were added manually.

[4] This is not the case at the European Parliament where interpreting into B or retour has gained more ground only after the enlargement of the Union. Virtually all the interpreters in EPIC only work into their native language.

[5] Each conference participant's speech event was assigned a progressive number throughout the transcript of a whole conference session (this system was inspired by a similar feature found in the EP verbatim reports).

[6] Lemmatization is particularly important when working on highly inflected languages such as Italian and Spanish.

[7] See <http://www.transana.org/>.

[8] For a thorough discussion of different display options of transcripts and software tools see Niemants (2012) and Setton (2011).

[9] See < http://www.sslmitdev-online.sslmit.unibo.it/corpora/corporaproject.php?path=E.P.I.C.>.

[10] See <http://catalog.elra.info/product_info.php?products_id=1145>.

[11] See <http://cartago.lllf.uam.es/static/dir-si/dir-si.html>.

REFERENCES

Baker, M. (1993) "Corpus linguistics and translation studies: Implications and applications". In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 233-250.

Bani, S. (2016) "Interpreting journalism". In C. Bendazzoli & C. Monacelli (eds.) *Addressing Methodological Challenges in Interpreting Studies Research*. Newcastle upon Tyne: Cambridge Scholars Publishing, 173-196.

Bendazzoli, C. & Sandrelli, A. (2009) "Corpus-based interpreting studies: Early work and future prospects". *Tradumatica 7. L'aplicació dels corpus linguistics a la traducció*. Online: <http://webs2002.uab.es/tradumatica/revista/num7/articles/08/08art.htm> (accessed 15 January 2016).

Bendazzoli, C. (2010) *Corpora e Interpretazione Simultanea*. Bologna: Asterisco. Open access: <http://amsacta.unibo.it/2897/> (in Italian, accessed 15 January 2016).

Bendazzoli, C. (2012) "From international conferences to machine-readable corpora and back: An ethnographic approach to simultaneous interpreter-mediated communicative events". In F. Straniero Sergio and C. Falbo (eds.) *Breaking Ground in Corpus-Based Interpreting Studies*. Frankfurt am Main [etc.]: Peter Lang, pp. 91-117.

Bendazzoli, C. (2016) "The ethnography of interpreter-mediated communication: Methodological challenges in fieldwork". In C. Bendazzoli and C. Monacelli (eds.) *Addressing Methodological Challenges in Interpreting Studies Research*. Newcastle upon Tyne: Cambridge Scholars Publishing, 3-30.

Bernardini, S. & Castagnoli, S. (2008) "Corpora for translator education and translation practice". In E. Yuste (ed.) *Topics in Language Resources for Translation and Localisation*. Amsterdam/Philadelphia: John Benjamins, 39-55.

Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bowker, L. & Pearson, J. (2002) *Working with Specialized Language. A practical Guide to Using Corpora*. London/New York: Routledge.

Brown, G. & Yule, G. (1983) *Discourse Analysis*. Cambridge: Cambridge University Press.

Christ, O. (1994) "A Modular and Flexible Architecture for an Integrated Corpus Query System". In *Proceedings of COMPLEX'94 3rd Conference on Computational Lexicography and Text Research Budapest, Hungary, July 7-10 1994*, 23-32. Online: <http://cwb.sourceforge.net/files/Christ1994.pdf> (accessed 15 January 2016).

Dembry, C. & Love, R. (2015) "Collecting the spoken BNC2014 – overview of methodology". Paper presented at the Corpus Linguistics 2015 Conference. Lancaster University, UK.

July 2015.

European Parliament (2014) “Rules of procedure”. 8th Parliamentary Term. Online: <http://www.europarl.europa.eu/sipade/rulesleg8/Rulesleg8.EN.pdf> (accessed 15 January 2016).

Fantinuoli, C. & Zanettin, F. (eds.) (2015) *New Directions in Corpus-based Translation Studies*. Berlin: Language Science Press.

Gile, Daniel (1994) “Opening up interpretation studies”. In M. Snell-Hornby, F. Pöchhacker and K. Kaindl (eds.) *Translation Studies: An Interdiscipline*. Amsterdam/Philadelphia: John Benjamins, 149-158.

Hayashi, R. (1996) *Cognition, Empathy, and Interaction: Floor Management of English and Japanese Conversation*. Norwood NJ: Ablex.

Heritage, John (1995) “Conversation analysis: Methodological aspects”. In U. M. Quasthoff (ed.) *Aspects of Oral Communication*. Berlin/New York: Walter de Gruyter, 391-416.

Hu, K. & Tao, Q. (2013) “The Chinese-English conference interpreting corpus: Uses and limitations”. *Meta* 58(3): 626-642.

Laviosa, S. (2011) “Corpus-based translation studies: Where does it come from? Where is it going?”. In A. Kruger, K. Wallmach and J. Munday (eds.) *Corpus-based Translation Studies. Research and Applications*. London/New York: Continuum, 13-32.

Leech, G. (1997) “Introducing corpus annotation”. In R. Garside, G. Leech and A. Mc Enery (eds.) *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London/New York: Longman, 1-18.

Metzger, M. & Roy, C. (2011) “The first three years of a three-year grant. When a research plan doesn’t go as planned”. In B. Nicodemus and L. Swabey (eds.) *Advances in Interpreting Research: Inquiry in Action*. Amsterdam/Philadelphia: John Benjamins, 59-84.

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). Online: http://www.erudit.org/revue/meta/2005/v50/n4/019850ar.pdf (accessed 15 January 2016).

Niemants, N.S.A. (2012) “The transcription of interpreting data”. *Interpreting* 14(2): 165-191.

Oléron, P. & Nanpon, H. (1965) “Recherches sur la traduction simultanée”. *Journal de Psychologie Normale et Pathologique* 62: 73-94.

Pöchhacker, Franz (1994) *Simultandolmetschen als komplexes Handeln*. Tübingen: Gunter Narr.

Russo, M. (2010) “Reflecting on interpreting practice: Graduation theses based on the European Parliament Interpreting Corpus (EPIC)”. In L.N. Zybatow (ed.)

Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungen zur Translationswissenschaft VI. Frankfurt: Peter Lang, 35–50.

Russo, M., Bendazzoli, C., Sandrelli, A. & Spinolo, N. (2012) "The European Parliament Interpreting Corpus (EPIC): implementation and developments". In Straniero Sergio, F. and C. Falbo (eds.) *Breaking Ground in Corpus-Based Interpreting Studies*. Frankfurt am Main [etc.]: Peter Lang, pp. 53-90.

Sacks, H., Schegloff, A. E. & Jefferson, G. (1974) "A simplest systematics for the organization of turn-taking for conversation". *Language* 50(4): 696-735. Online: <http://www.jstor.org/stable/412243> (accessed 15 January 2016).

Sandrelli, A. (2010) "Corpus-based Interpreting Studies and interpreter training: A modest proposal". In L. N. Zybatow (ed.) *Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungen zur Translationswissenschaft VI*. Frankfurt: Peter Lang, 69–90.

Schmid, H. (1994) "Probabilistic part-of-speech tagging using decision trees". *Proceedings of International Conference on New Methods in Language Processing. September 1994*. Online: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf> (accessed 15 January 2016).

Schmid, H. (1995) "Improvements in part-of-speech tagging with an application to German". *Proceedings of the ACL SIGDAT-Workshop. March 1995*. Online: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf> (accessed 15 January 2016).

Setton, R. (2011) "Corpus-based Interpreting Studies (CIS): Overview and prospects". In A. Kruger, K. Wallmach and J. Munday (eds.) *Corpus-based Translation Studies. Research and Applications*. London/New York: Continuum, 33-75.

Shlesinger, M. (1998) "Corpus-based Interpreting Studies as an offshoot of Corpus-based Translation Studies". *Meta* 43(4): 486-493.

Thompson, P. (2005) "Spoken language corpora". In M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books: 59-70. Online: <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm> (accessed 15 January 2016).

Wang, B. (2012a) "A descriptive study of norms in interpreting: Based on the Chinese-English Consecutive Interpreting Corpus of Chinese Premier Press Conferences". *Meta* 57 (1): 198-212.

Wang, B. (2012b) "Interpreting strategies in real-life interpreting: Corpus-based description of seven professional interpreters' performance". *Translation Journal* 16 (2). Online: <http://translationjournal.net/journal/60interpreting.htm> (accessed 15 January 2016).

外研社 高等英语教育学术会议文集

口译研究：新视野　新跨越
——第十届全国口译大会暨国际研讨会论文集

“全国口译大会暨国际研讨会”自1996年首次召开以来，每两年召开一次，开创了全国性口译学术研讨的先河。此后，中国口译学人两年相聚一次，在讨论中碰撞思维，在交流中收获成长，共同见证了中国口译教育和研究的进步和中国翻译事业的发展。

本书是2014年“第十届全国口译大会暨国际研讨会”论文集，收入论文及发言稿共24篇。

全书分三个部分：

- 口译教学研究：收录10篇论文，从宏观和微观层面探讨了口译教育、口译教学和口译课堂的方方面面。
- 口译理论研究：收录8篇论文，围绕口译语料库、口译能力、译员角色和责任、口译测试、译文质量等主题展开研究讨论。
- 专家论坛：收录6篇发言稿，真实地再现了大会的主旨演讲环节部分内容。

高等英语教育出版分社宗旨：
推动科研·服务教学·坚持创新
外研社·高等英语教育出版分社
FLTRP Higher English Education Publishing
电话：010-88819595
传真：010-88819400
E-mail：ced@fltrp.com
网址：http://www.heep.cn
（教育网 http://edu.heep.cn）

记载人类文明
沟通世界文化
www.fltrp.com

责任编辑：卫　昱
封面设计：郭　莹
版式设计：涂　俐

ISBN 978-7-5135-8150-9
9 787513 581509 >
定价：98.00元

口译研究：新视野　新跨越——第十届全国口译大会暨国际研讨会论文集

外研社

外研社 高等英语教育学术会议文集

口译研究：新视野　新跨越
——第十届全国口译大会暨国际研讨会论文集

Interpreting Studies: The Way Forward
—Proceedings of the 10th National Conference and International Forum on Interpreting

主　编：陈　菁　杨柳燕

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

序　言

1996年，首届全国口译大会在厦门大学召开，开创了全国性口译学术研讨的先河。此后，中国口译学人两年一次定期相聚，在讨论中碰撞思维，在交流中收获成长，共同见证了中国口译教育和研究的进步和中国翻译事业的发展。作为我国学术水平最高、影响力最广泛的口译学术盛会，全国口译大会有力地推动了我国口译教学与研究的发展，加强了我国口译界学者和从业者之间以及与国际同行间的学术交流。2014年10月，来自海内外的312位口译教育者和研究者们再一次齐聚厦门，一同迎来全国口译大会的十周岁生日。

第十届大会由中国译协和厦门大学联合主办，厦门大学外文学院、厦门大学口译学研究所承办。大会参会代表来自中国大陆、香港特区、台湾地区、澳门特区、澳大利亚、美国、英国、奥地利、意大利等10余个国家和地区，参会论文200余篇。大会围绕“口译研究：新视野 新跨越”的主题，以主旨发言、主旨论坛、分组论坛和论文海报展示等形式，针对跨学科口译研究、口译教育、口译测试、手语传译、语料库口译研究和口译服务等议题展开了热烈和卓有成效的研讨。大会同期举办了第五届海峡两岸口译大赛总决赛、“我与博导有约”学术沙龙、中国手语文化之夜等丰富多彩的配套活动。大会首次将中国手语与汉语和英语一起列为会议的工作语言，聋哑人译员第一次走上了口译大会的舞台并与听人译员搭档进行了高水准的同传协作，聋哑人学者也参与会议并用手语进行讨论和发言。大会首次设立论文海报展示，为参会代表开辟了崭新的论文交流方式。

本届大会历时两天。在10月17日的开幕式上，中国译协常务副会长唐闻生女士、厦门大学詹心丽副校长、广东外语外贸大学仲伟合校长、欧盟口译总司口译部主任布莱恩·福克斯（Brian Fox）先生、厦门大学外文学院张龙海院长分别致辞，表达了对大会召开的热烈祝贺和美好祝愿。在主旨演讲环节，北京语言大学高翻学院刘和平教授、美国加劳德特大学辛西娅·罗伊（Cynthia Roy）教授、广东外语外贸大学董燕萍教授、英国赫瑞瓦特大学克劳迪娅·安吉莱利（Claudia Angelelli）教授携手厦门大学陈菁教授、意大利都灵大学科罗迪·本德左立（Claudio Bendazzoli）博士分别就中国口译教育、话语分析和口译教学、语言使用生态与认知控制、口译

职业资格认证测试以及语料库口译研究等话题发表了他们的真知灼见。

大会的第一天下午和第二天上午是分论坛研讨环节。在以口译教育（72篇论文）、跨学科口译研究（26篇论文）、语料库口译研究（13篇论文）、手语翻译（12篇论文）、口译测试（12篇论文）、口译策略、技术和服务（58篇）为主题划分的分论坛中，与会代表展开了积极的思想碰撞。

大会的第二天下午包括两个环节：分论坛汇报和主旨论坛。首先5位分论坛组长向大会汇报各组研讨的基本情况和热点话题。紧接着，一场题为“口译研究：新视野 新跨越”的主旨论坛拉开了序幕。论坛由维也纳大学弗朗茨·彼赫哈克（Franz Pöchhacker）博士主持，仲伟合教授、布莱恩·福克斯先生、北京外国语大学高翻学院院长王立弟教授、上海外国语大学高翻学院名誉院长柴明颎教授和台湾辅仁大学跨文化研究所所长杨承淑教授等知名专家担任论坛嘉宾。他们从国家政策、口译教育、口译研究、口译服务、口译市场等角度描绘了口译研究的宏图，展望了口译的未来，将本届大会的学术研讨推向了高潮。在简短的闭幕式环节，全体与会者在一部由厦大口译团队精心制作的“译言心声”的短片中忆往昔、展未来，感慨中国口译事业的光阴流转和春华秋实。大会在感动的泪水、期待的眼神和热烈的掌声中落下了帷幕。

在本届大会中，口译研究的新视野和新跨越得以彰显。参会者从语言学、社会学、心理学、认知科学、传播学等诸多视角探讨了口译研究的过去、现在和未来，展现了丰富的研究视角和多元的研究手段。大会着力探讨了口译研究的跨学科途径，凸显了跨学科研究的意识、思路、手段和方法在口译研究中的核心作用。以主旨演讲为例，大会一方面鼓励演讲者们从不同学科视角探讨口译，如辛西娅·罗伊教授从语篇分析、克劳迪娅·安吉莱利和陈菁教授从语言测试学、科罗迪·本德左立从语料库等视角寻求口译学与其他学科的融合，另一方面还专门邀请了心理语言学和应用语言学方向的董艳萍教授做了题为“语言使用生态与认知控制优势：双语学习、演讲及口译训练”的演讲。在分组研讨中，80%以上的发言采用了跨学科研究途径，他们或是借鉴了其他学科的术语、概念和理论体系，或是运用了其他学科的方法与工具。这些跨学科视角对丰富口译学研究思路、改善研究方法、推动学科发展起着不可估量的作用。本届大会致力于为年轻学者提供自由发表学术思想和引领学术讨论的平台。大会大胆启用年轻学者担任分论坛主持人，点评和总结发言，同时在“我与博导有约”的学术沙龙中突出了年轻学者的主角地位，以鼓励学术独立、培养学术自信。

本册论文集是在对近百篇所提交的论文全文进行评审遴选后集结而成的。论文

分三个部分：口译教学研究、口译理论研究和专家论坛。

在“口译教学研究”部分，12位作者从宏观和微观层面探讨了口译教育、口译教学和口译课堂的方方面面；

在“口译理论研究”部分，10位作者围绕口译语料库、口译能力、译员角色和责任、口译测试、译文质量等主题展开研究；

本论文集的“专家论坛”部分真实地再现了大会的主旨演讲环节：弗朗茨·彼赫哈克博士在这部分的开篇之作中介绍了五位主旨论坛嘉宾以及他们的发言，接着勾勒了口译研究的发展脉络和未来路向。仲伟合教授从学科界定出发，探讨了口译的学科建设状态及指标与学科定位，从口译研究、人才培养及实践服务三个方面总结并分析了口译学科的发展现状与存在问题，提出了我国口译学科未来发展的六个纬度。布莱恩·福克斯先生阐述了欧盟使用口译的过去和未来，展望了技术带来的挑战和机遇。王立弟教授从传统和创新角度入手，论述了口译训练中的技能培养和译者培养的核心要素。柴明颎教授介绍了口笔译训练与学术研究领域的国际合作情况。杨承淑教授介绍了台湾辅仁大学推出的一个跨领域、跨媒体、跨国界的国际医疗口笔译硕士学分学程，展示了口译教育的新思路。

虽然受篇幅所限，本论文集所收录的论文数量有限，但从中不难看出，口译研究视域更加宽广，理论思考更趋深入，方法选择更显合理。内容上，除了口译教学之外，研究者们还热衷于探讨口译过程、口译产品、资格认证、职业伦理、手语传译、语料库研究等话题。视角上，跨学科借鉴成为当前口译研究的主流思维。研究者不仅从语言学、传播学、跨文化研究等相邻学科获得启示，还从心理学、认知科学、神经科学等前沿领域汲取养分。方法上，实证主义思维深入人心，研究者充分运用实验法、问卷调查、现场观察、当面访谈、口译语料库等多种途径搜集和分析数据，极大地丰富了口译研究的工具手段。所有这些都预示着中国的口译研究正在步入一个多元健康的发展阶段。

陈　菁　杨柳燕

2016年5月1日于厦大芙蓉园

目　录

第一部分　口译教学研究

第二部分　口译理论研究

第三部分 专家论坛