Introduction

This chapter gives an overview of corpus-based methods and studies in conference interpreting. The content is organised into five main sections: methodology; overview of conference interpreting corpora; research deliverables; corpus applications to interpreting education and professional practice; and future prospects.

The methodological section introduces the corpus-based approach and explains how conference interpreting corpora are created. The fundamental steps of corpus design, data collection, transcription, annotation, access and distribution are illustrated with reference to major corpus projects.

Against this background, the second section gives an overview of existing conference interpreting corpora. They are grouped according to types of setting and communicative situation, since these cut across language combinations and may entail different data collection strategies. In this respect, the reason why some particular settings have lent themselves to scholarly attention more than others lies in the availability and accessibility of data. The settings studied to date thus range from international institutions (e.g. the European Parliament with simultaneous interpreting) to governmental press conferences in China (with consecutive interpreting) and local markets with a variety of communicative situations.

Turning to research deliverables, the next section reports on the main results and observations from conference interpreting corpora. Here, looking at the selected studies can obviously not provide an exhaustive picture of what is known about conference interpreting thanks to corpora. Nevertheless, even a brief survey of this kind affords useful insight into the research potential of these language resources, helping to contextualise corpus-based investigation as compared with other approaches to interpreting research.

The fourth section examines the role of corpora in conference interpreting education and professional practice, two areas where the potential contribution of corpus-based studies is still underexplored.

Looking at future prospects, the final section outlines a number of possible developments related to corpus features, research areas, and applications.
Developing conference interpreting corpora

The corpus-based approach was introduced in interpreting studies towards the end of the 1990s (Shlesinger 1998), in the wake of the pioneering work done in translation research (Baker 1993). The availability of large datasets of written texts in electronic form made it possible to use computer programs to extract and count occurrences of various kinds of phenomena, accounting for typical and untypical features of translated texts. In interpreting studies, the same methodology was proposed as an attempt to overcome the limitations posed by “sparse, often anecdotal data” (Shlesinger 1998: 2) and to scrutinise larger samples of authentic data. A fundamental difference with respect to translation corpora is that interpreting corpus data need to be transcribed. As illustrated below, transcription entails critical methodological choices and is still considered a time-consuming activity, despite the availability of speech recognition programs.

A corpus can be defined as “a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria” (Bowker & Pearson 2002: 9). The methodological steps in corpus development are corpus design, data collection, transcription, annotation, alignment, data access and distribution (Bendazzoli 2010; Bernardini et al. 2018). While the first three steps are essential (as there can be no corpus without transcribed data), the remaining steps are optional but contribute substantially to increasing the research potential of an interpreting corpus.

Looking at the development of corpus-based interpreting studies, it is possible to find two major features that often set interpreting corpora apart from the definition quoted above (Bendazzoli & Sandrelli 2009). First, some interpreting corpora are not electronic. In these cases, the term corpus is used for data samples that are analysed manually, without the aid of corpus linguistics methods, and are of limited size. Second, not all interpreting corpora are openly accessible—e.g. on internet-based platforms. In addition, early studies conducted with the corpus-based approach concerned only conference interpreting, particularly in the simultaneous mode. Over time, the approach has been extended to other modes and settings, not necessarily limited to conference interpreting (Bendazzoli 2018; Setton 2011).

The limitations mentioned above are related to certain special challenges in the development of interpreting corpora. For conference interpreting corpora in particular, each stage of corpus development involves a number of choices that can, in turn, usher in potential constraints. In corpus design it is necessary to consider what kind of interpreting mode will be represented in the corpus—whether simultaneous (with or without a booth, whispered interpreting, sight translation, see Bartłomiejczyk & Stachowiak-Szymczak, Chapter 2, in this volume), consecutive, or both; how many working languages are involved; and the desired level of representativeness (Halverson 1998). It is possible to organise the data into parallel corpora, including source speeches and the resulting target speeches; comparable corpora, including source and target speeches in the same language; and intermodal corpora, including multiple target texts obtained from the same source but with different translation modes—e.g. simultaneous interpreting and written translation (Bernardini et al. 2018; Shlesinger 2008).

Interpreting corpora normally rely on transcripts of recorded data—i.e. spoken and/or sign language data. Technological advances have made it easier to video record communicative situations, though several projects have been based on audio recordings only. In either case, data collection can be accomplished directly through fieldwork (Bendazzoli 2016), or indirectly—by such means as downloading data from the internet (e.g. the European Parliament video library) or recording interpreter-mediated events broadcast on TV. Fieldwork-based data collection entails different (logistic and technical) challenges, depending on the interpreting...
modes and settings under consideration. For instance, in consecutive interpreting it would be interesting to keep track of the interpreter’s notes along with the production of source and target speeches. This is now possible with digital pen technology (Kellet Bidoli 2016; see also Ahrens & Orlando, Chapter 3, in this volume). In simultaneous interpreting without a booth and in whispered interpreting, special care must be given to the recording equipment in order to obtain high quality recordings of source and target speeches (e.g. by means of Lavalier microphones). In simultaneous interpreting with a booth, it is sometimes necessary to coordinate data collection from different locations, as the booth may not be installed in the same room where the conference proceedings take place. Furthermore, in simultaneous interpreting, there are (at least) two simultaneous audio tracks to be managed at the same time, and these should be kept separate so as to transcribe them conveniently.

The next step involves transcription, which can be managed with different software programs (Niemants 2012) and conventions, depending on what is to be the focus of the analysis and on the interaction format (i.e. monologue vs. dialogue). Transcripts may undergo annotation of certain attributes, again depending on the analysis to be carried out. Generally, metadata annotations are recorded to keep track of the kind of text/speech one is dealing with, including features such as the total number of words, the speed of delivery, its duration, speaker-related details, and so on. In addition, further (optional) levels of annotation range from part-of-speech (POS) tagging—i.e. the annotation of grammatical categories for each word (or token)—to specific verbal or nonverbal features (e.g. mispronounced words). Annotations can be applied automatically (as in the case of POS-tagging) or manually. In sign language interpreting (see Turner, Grbić, Stone, Tester & de Wit, Chapter 38, in this volume), both the transcription and the annotation processes undergo further challenges: the need to use glosses of signs; the variation resulting from the three-dimensional nature of signs themselves; and the impossibility of using transcripts as separate files if they are embedded in a tier of ELAN, the transcription program widely used in sign language research (Wehrmeyer 2015).

Transcripts may also be processed for alignment, which in interpreting corpora can be set up on two levels. First, there can be alignment between source speech and target speech transcripts for immediate comparison (Volk 2019). Depending on the interpreting mode and on corpus design (intermodal corpora entail a more complex structure, inter alia in terms of alignment—see Bernardini et al. 2016), aligned transcripts can be displayed in tabular form or following a musical score layout. In addition, there can also be alignment between a transcript and its video/audio recording (usually time codes are annotated into the transcript).

Finally, the last step in corpus development concerns its accessibility and distribution. Some corpora are made available to the research community with no particular restrictions, whereas others are only accessible to the researchers who developed them (e.g. for confidentiality issues, or by copyright). Ready-made corpora are extremely useful language resources for replications of past studies or for new inquiries. They can also be used to develop further corpora, with additional language combinations or specific purposes in mind (Bendazzoli et al. 2021). Conversely, DIY (i.e. do-it-yourself) corpora require considerable effort and are more likely to be created within larger team-based research or for PhD projects.

**Overview of conference interpreting corpora**

Corpus research has been especially productive in conference interpreting settings since the early developments of this research paradigm in interpreting studies. This is probably related to two factors: the longer research tradition centred on conference interpreting than on other types of interpreting—some of the pioneers of early interpreting corpora came from a conference
interpreting background; and, arguably even more relevant, the relatively greater accessibility of conference interpreting settings. The European Parliament (EP), along with other EU and international institutions, is a case in point. Owing to the principle of transparency, which is applied by many international institutions (now granting direct access online to interpreted events), and to the extensive use of professional interpreting services, EP plenary sittings have lent themselves to the creation of several corpus projects.

Some of these investigations are based on transcripts of source and target speeches and may even feature text-video alignment, but are still analysed manually (e.g. Vuorikoski 2004). Other studies leverage automatic queries to detect occurrences of certain features in verbatim reports, which are then checked against delivery and discussed more qualitatively (e.g. Bartłomiejczyk 2020).

The first project to deliver a fully accessible, machine-readable corpus of EP interpreting was carried out at the University of Bologna at Forlì (Italy) and resulted in the European Parliament Interpreting Corpus (EPIC) (Monti et al. 2005). EPIC consists of nine sub-corpora of transcribed and annotated speeches: three sub-corpora of source speeches in Italian, English, and Spanish, and six sub-corpora of simultaneously interpreted speeches covering all the possible combinations among the three languages involved. The particular structure of EPIC makes it possible to use it both as a parallel and as a comparable corpus. The EPIC project has also been the springboard for other similar corpora including different language combinations, such as English, French, Dutch and (only as a source language) Spanish in EPIC-G (EPIC-Ghent).

Given the unique features of the EP setting, the transcripts of oral speeches and simultaneous interpretations have also been analysed against the relevant written proceedings and their translation, thus taking an intermodal perspective (Defrancq et al. 2015). Along the same lines, EPIC has also been turned to be intermodal, covering other language combinations (i.e. English, French, Italian, along with some sub-corpora of Polish, Slovene and Finnish) and data from different EP sessions, which has resulted in the European Parliament Translation and Interpreting Corpus (EPTIC) (Bernardini et al. 2016). On the other hand, an example of how comparable corpora can be developed from similar recorded material is the Translation and Interpreting Corpus (TIC), comprising EP source speeches and texts in English along with simultaneously interpreted speeches and translated texts into English from French, Spanish, German and Dutch (Kajzer-Wietrzny 2012). Another example of the successful exploitation of a ‘ready-made’ corpus such as EPIC is a purpose-specific corpus called Anglintrad (Bertozzi 2018), which was developed to look at how Anglicisms are managed in interpreting and translation from Italian into Spanish. The aim of this corpus is primarily to inform translator and interpreter education (see next section).

Other than the EP, there are fewer corpora based on data from other EU institutions. For instance, the IMITES corpus (Spinolo 2018) comprises Italian/Spanish data from the European Commission to study figurative language in simultaneous interpreting.

Similarly favourable conditions for data collection can be found in the context of the United Nations (see Ruiz Rosendo & Diur, Chapter 9, in this volume), which provides access to the meetings of the General Assembly and other interpreter-mediated events through its online TV channel. However, the development of conference interpreting corpora based on this specific setting is still quite limited: the SIREN corpus (Dayter 2018) is seemingly the first attempt to build an electronic corpus of simultaneous interpretations between Russian and English.

TV streaming online or even traditional TV broadcasts are another fruitful source of data for conference interpreting corpora. Fifty years of televised interpreter-mediated events have been collected in the Italian Television Interpreting Corpus (CorIT) (Straniero Sergio 2007), which includes consecutive and simultaneous interpretations into Italian coming from a variety of
events broadcast on Italian TV channels (Dal Fovo 2013). Not all the target speeches contained in this corpus are paired with the relevant source speeches, but research based on this particular resource has proved useful to pinpoint special features of interpreting on television (Falbo 2012) and of interpreters’ style (Straniero Sergio 2012).

A more targeted set of data obtained from TV broadcasts makes up the bilingual parallel corpus of Chinese-English Interpreting for Premier Press Conferences (CEIPPC), which was used to investigate consecutive interpreting norms (Wang 2012) and language-specific features such as syntactic asymmetry between Chinese and English (Wang & Zou 2018). China’s premier press conferences (see also Dawrant, Wang & Jiang, Chapter 15, in this volume) have also been considered in other corpora. Examples are the Chinese-English conference interpreting corpus, or CECIC (Hu & Tao 2013), with the addition of written reports and a comparable dataset from Web-sourced CNN press conferences; and another corpus including simultaneous interpreting and translated texts (Fu 2016). Further investigations have taken advantage of the availability of interpreting data (obtained from TV or online) relating to political speeches at high-profile events held in China and other parts of Asia. For instance, a critical discourse analysis approach was adopted to investigate ideology and self-referentiality in the Chinese-English Political Discourses Corpus (CE-PolitDisCorp) (Gu & Tipton 2020). Pragmatic markers were analysed in the Corpus of Interpreted Political Speeches from Chinese to English (CIPSCE) (Pan & Wong 2019), including both consecutive and simultaneous interpreting. Data of this kind, along with translated texts, have been gathered to create a new intermodal corpus similar to EPTIC—the Chinese/English Political Interpreting Corpus (CEPIC) (Pan 2019). CEPIC is a large, machine-readable corpus of conference interpreting and translation of political speeches. It is also accessible online, through a dedicated platform. At the time of writing, this resource is seemingly the largest conference interpreting corpus of professional Chinese (Cantonese and Putonghua)-English interpreting and translation.

Conference interpreting on television can also provide other research opportunities, especially in the case of special events with global coverage and media services. For instance, taking an intermodal perspective, Taehyung (2011) analysed English-Korean simultaneous interpreting vs. live captions in three different editions of the Academy Awards Ceremony. In addition, the presence of sign language interpreters as an increasingly common feature of many daily news services and live addresses—e.g. during natural disasters and pandemics—affords an opportunity to create sign language interpreting corpora (Wehrmeyer 2019).

Direct access to conference interpreting data has been managed in various ways. The CoSi corpus (Meyer 2008) includes Portuguese-German interpretations of the same talk held in three different venues, two with consecutive and one with twin simultaneous interpreting booths. The researchers had full access to this real-life event, as they were associated with the organisation of the conference in a supporting capacity. The CoSi corpus is freely available on the EXMARaLDA portal (House et al. 2012). Another data collection strategy entails the direct involvement of a practisearcher—i.e. a professional interpreter (on duty) who is also a scholar engaged in research (see also Gile & Barranco-Droege, Chapter 25, in this volume). Examples range from international medical conferences, as in the DIRSI Corpus (Bendazzoli 2012), to pre- and post-football match press conferences in the FOOTIE Corpus (Sandrelli 2012, see also Sandrelli, Chapter 6, in this volume).

Finally, there are conference interpreting corpora compiled by researchers who apparently acted as external observers or obtained recordings from the organisers, as is the case with the WAW corpus, which includes English-Arabic/Arabic-English simultaneous interpretations of conference presentations from three different events held in Qatar, totalling 521 sessions and 119 hours (Abdelali et al. 2018).
Claudio Bendazzoli

This overview of conference interpreting corpora highlights that there are various data sources, with some settings proving more accessible than others. It can also be seen that conference interpreting corpora are becoming larger and increasingly machine-readable, though smaller datasets can still be scrutinised with traditional methods without the aid of computer programs. In this respect, small corpora are ideal for qualitative approaches (e.g. Petite 2005), while quantitative results can set the basis for in-depth, qualitative investigations (e.g. Bendazzoli 2019).

What we know about conference interpreting thanks to corpora

Considering the relatively limited size of many interpreting corpora (especially in comparison with monolingual reference corpora), the results reported so far can hardly be generalised. In addition, the numerous variables involved further narrow down the scope of many studies, as their focus differs in terms of language combinations, language directions, interpreting modes, directionality and communicative situations. It can be argued that EP plenary sittings are the most extensively scrutinised conference interpreter-mediated situation, followed by political discourse in China’s governmental press conferences. In terms of languages, English is the most represented, as a source and target language alike. This is not surprising, given the extensive use of English, both as a native and a non-native language (see Albl-Mikasa, Chapter 39, in this volume) in most settings, and the early development of corpus linguistics tools able to process English transcripts (e.g. automatic taggers).

As mentioned above, corpus research into conference interpreting has been carried out both manually, thus looking at the data without the aid of any particular computer software, and automatically—i.e. extracting occurrences by means of IT tools and, in some cases, applying statistical testing and measurements. The latter approach might be perceived with scepticism by the more sociologically-minded scholar, as an excessively dry take based on mere number crunching and thus devoid of situatedness. However, quantitative studies have also opened the way to more in-depth, qualitative investigations, which are in turn corroborated by well-founded results and in no way limited to sporadic or impressionistic evidence.

Overall, attention has been focused on several areas of interest, such as verbal and non-verbal features of source and target speeches, manifestations of cognitive load and interpreting strategies (see Riccardi, Chapter 27, in this volume), differences between interpreting and translation of the same source text, the role of gender (see Defrancq, Collard, Magnifico & Iglesias Fernández, Chapter 30, in this volume), and ideology.

Regarding EP discourse, different corpus projects have revealed some salient features of source speeches—e.g. short duration in most speech events, high speed of delivery, and read or semi-prepared presentation style (de Manuel 2003b; Russo et al. 2012; Vuorikoski 2004). On the other hand, target speeches (i.e. interpreters’ output) have been analysed from different perspectives. Looking at EPIC as a comparable corpus, a study of lexical patterns revealed that lexical density is higher in interpreted speeches than original speeches (in the same language) (Russo et al. 2006). This result runs counter to findings of lexical simplification obtained in translation (Laviosa 1998), but is corroborated by the results obtained in the English component of the SIREN corpus (Dayter 2018). Conversely, with the sole exception of the Italian booth in EPIC, lexical variety was found to be lower in interpreted speeches, this trend being in line with Laviosa’s findings and with the Russian component of SIREN. The same result was also found in EPTIC, where target texts show less lexical variety than source texts and interpreted texts are ‘simpler’ than translated texts (Bernardini et al. 2016).
A comparable corpus was also used by Kajzer-Wietrzny (2018), to analyse linguistic patterns in interpreted English by comparison with native and non-native English. Focusing on the complementiser that, a higher degree of explicitation is found in interpreted English; and greater use of the optional that is found in both interpreted English and non-native English source speeches vs. native ones. Again from a comparable corpus perspective, pragmatic markers were analysed in the CIPSCE (Pan & Wong 2019) and, with the exception of contrastive and elaborative markers, were found to be underused in interpreted speeches.

Moving to nonverbal features, the presence of disfluencies in interpreters’ and source speakers’ output was investigated in EPIC (Bendazzoli et al. 2011) and also in the SIREN corpus (Dayter 2021). Despite the consideration of different sets of features (mispronounced and unfinished words in EPIC; mispronounced words, hesitations, self-repairs, silent pauses, repetitions and fillers in SIREN), the two studies found similar patterns to a limited extent, with more disfluencies in target speeches than original speeches and fewer repairs in interpreters’ output. However, unfinished words were more frequent in English source speeches than interpreted English in EPIC, pointing to the important role played by individuals’ speaking style (which can also be true for interpreters—see Kajzer-Wietrzny 2013).

Besides the EP setting, the speaking style of interpreters has been the object of analysis in other corpus studies, namely in TV interpreting through CORIT (Falbo 2012, see also Falbo, Chapter 7, in this volume). The systematic observation of the data included in CORIT has helped to outline the main characteristics of talk show interpreting (Straniero Sergio 2007) and reveal that individual interpreters tend to make recurrent use of a limited number of set-phrases and expressions, including additional discourse markers that do not correspond to equivalent units of meaning in the source speech. The occurrence of interpreter-generated discourse markers was also confirmed in studies based on other corpora. For instance, in the DIRSI corpus, 30 per cent of all the occurrences of the discourse marker so in target speeches are added by the interpreters (Bendazzoli 2019); in EPIC-G (Defrancq et al. 2015), discourse markers are omitted in interpretations but also added—often to a greater extent than in the case of translation. This intermodal perspective has shed light on different patterns found in translated texts from those found in interpreting. For example, in EPTIC, interpreted speeches show a higher degree of simplification (Bernardini et al. 2016; Ferraresi et al. 2018) and less frequent phraseological patterns, along with a greater use of non-standard word combinations, compared to translated texts and also to original texts (Ferraresi & Miličević 2017). Rather than contrasting interpreting with translation, the CoSi corpus was used intermodally to analyse interpreters’ intervention in the renditions of proper names in simultaneous and consecutive interpreting (Meyer 2008). While there is ample evidence of variation in the way proper names are conveyed in target speeches, again denoting individual interpreters’ style, no particular difference emerged between simultaneous and consecutive interpreting.

The analysis of textual and interactional features in source speeches and/or interpreters’ output has been extended to account for interpreting strategies, norms, and cognitive load. Simultaneous interpreting corpora of football press conferences (Sandrelli 2017) and presidential debates (Dal Fovo 2018) have yielded insightful results about interpreters’ management of interactional patterns, highlighting the occurrence of coordination activities even among simultaneous interpreters, along with differences in professional practice depending on team composition and working conditions (e.g. on-site vs. off-site interpreters).

As regards consecutive interpreting in a corpus of Chinese-English governmental press conferences (Wang 2012), interpreting shifts are identified in terms of additions, reductions, and corrections, with additions being far more frequent than the others among the five interpreters...
represented in the corpus. Another example of tactics in consecutive interpreting is the study of language-specific features such as syntactic asymmetry between Chinese and English (Wang & Zou 2018), which is likely to have an impact on interpreters’ cognitive load as they transform long and complex front-loaded attributive modifying structures into back-loaded ones. Similar results have been obtained in another study of markers and predictors of cognitive load in simultaneous interpreting into Dutch, based on a comparable corpus perspective (Defrancq & Plevoets 2018; Plevoets & Defrancq 2018). Filled pauses were more frequent in interpreted Dutch, especially when interpreters were dealing with linguistic asymmetry, compound nouns and higher lexical density, whereas formulaic patterns were inversely correlated to this manifestation of cognitive load.

Two further areas of scrutiny are gender and ideology. These lines of research have given rise to interesting expansions in the scope of corpus-based interpreting studies, clearly showing that quantitative data along with the systematic retrieval of occurrences can potentially afford an interesting starting point for qualitative, in-depth investigations. For instance, Russo (2018) found some statistically significant speaking patterns in a sub-section of EPIC, identifying higher speeds of delivery and longer target speeches by female interpreters than by their male colleagues. Similarly, Magnifico and Defrancq (2017) found evidence of gender-based patterns in interpreters’ use of hedges (in a sub-section of EPIC-G), with women hedging more than men. Opposite findings emerged in the management of politeness, with male interpreters toning down face-threatening acts more than female interpreters (Magnifico & Defrancq 2016; see also Defrancq, Collard, Magnifico & Iglesias Fernández, Chapter 30, in this volume).

The final area of research encompasses ideology and (critical) discourse analysis (see Okoniewska & Wang, Chapter 31, in this volume). Examples are again based on EP data (Bartłomiejczyk 2020; Beaton 2013; Vuorikoski 2004) or Chinese governmental press conferences (Gu & Tipton 2020), which are analysed either manually or automatically. These studies demonstrate how important it is for source speakers to be aware of the interpreting process through which the sense of their message is conveyed to target language listeners. They also reveal that multilingual environments, especially high-level conferences, pose special challenges to interpreters, from rhetorical strategies (such as self-referentiality) to irony, or from racist speech to management of conflicting views. While such features are clearly recognised, there seems to be a lack of commonly shared interpreting standards or protocols to deal with them.

**Corpora in conference interpreting education and professional practice**

The constant development of the corpus-based approach and the availability of corpora are having a slow, yet growing impact on conference interpreting education and, to a much lesser extent, professional practice. Easier access to conference interpreting data and the implementation of extensive conference interpreting projects have given many students the opportunity to take advantage of existing corpora for their final dissertations (Dal Fovo 2011; Russo 2010). EU institutions, especially the European Parliament, are once again the main data source for corpus-related pedagogical applications. Going back in time, Dollerup and Ceelen (1996) provide possibly one of the earliest examples of pedagogical corpora for consecutive interpreting training. Their corpus of consecutive interpreting in Danish, Dutch, English, French, German and Italian is in fact a collection of audiotaped recordings and transcripts, and not a machine-readable corpus proper, of lectures and trainees’ output within a training course held in 1976. A more advanced, digital development of this kind is the Marius database, a collection of speeches from the EP and the World Social Forum set up by de Manuel (2003a, 2003b) and
Corpus studies in conference interpreting

Corpus studies in conference interpreting have led to the development of pedagogical resources. For simultaneous interpreting, Sandrelli (2010) designed a pedagogical activity based on use of corpus tools to extract occurrences of verb tenses in Italian, English, and Spanish from EPIC and see how these are used by EP interpreters. Along the same lines, Aston (2018) argues that the analysis of concordances from a corpus of transcripts of EP interpreters’ output into English can be useful to enhance formulaic phraseologies among interpreters in autonomous learning activities. Spinolo (2018) organised training sessions on interpretation of figurative language using the IMITES corpus, while Bertozi’s (2018) Anglintrad corpus and platform afford a wealth of terminological information concerning the treatment of Anglicisms in interpreting and translating from Italian into Spanish. Cresswell (2018) combined different SI corpora to study the frequency of phrasal verbs in English target speeches and found a higher frequency in the output of native English interpreters vs. interpreters working into English as their B language (i.e., active foreign language). This finding, along with the retrieval of formulaic language from the corpora taken into account, prompted the development of targeted teaching material for conference interpreters, particularly in language enhancement modules. The analysis of collocations from an intermodal perspective—i.e., as found in translated and interpreted target texts from the same source text in EPTIC—has also led to the design of task-based activities in the translation classroom: conference interpreting data are contrasted with translation data and stimulate reflection on the differences between the two processes (Ferraresi 2016).

Non-interpreting corpora have also been exploited for pedagogical purposes in interpreter education. For example, Bale (2013) took advantage of the BACKBONE corpus (Kohn 2012) to improve trainees’ lexical knowledge as part of the language skills deployed in (English/German) consecutive interpreting. BACKBONE, a corpus of interviews in several languages, is unrelated to translation or interpreting per se. The usefulness of this corpus in interpreter training stems from its pedagogically-oriented annotations, which make it possible to select transcripts and videos according to language, topic, and particular language features. Interpreter trainees can thus use these materials as source speeches for interpreting practice, and as parallel texts for terminology building and checking. A similar implementation was proposed by Leeson (2008), for sign language interpreter training, with the Signs of Ireland corpus. This corpus can be used not only for retrieving source speeches, but also as a reference corpus to check and learn collocations, with scope for self-analysis of one’s interpreting performance. The last two examples clearly illustrate the educational potential of corpora, even when they do not represent mediated communication but have been complemented by pedagogically-oriented annotations.

As regards professional practice, it is a fair assessment to state that corpus methods have not yet broken through. While it is hard to see appreciable added value of corpora (as opposed to consultation of a good glossary) during an interpreting assignment, they may find applications prior to assignments for preparation and, more generally, to further terminological competence.
and language skills (e.g. with collocation searches and phraseology). In this respect, potential solutions may be found in web-crawled corpora (Baroni & Ueyama 2006), as source material from which to build ad-hoc comparable corpora and term banks (Fantinuoli 2018), as well as software programs. A case in point is InterpretBank (Fantinuoli 2016), which was specifically designed as a kind of workstation for interpreters, including tools for automatically extracting texts and terms, managing glossaries, and accessing terminology on the fly (see also Fantinuoli, Chapter 36, in this volume). Despite the obvious relevance of such resources to professional needs, the real impact of corpus development and research on conference interpreting practice has yet to be investigated in detail, especially among younger conference interpreters who may have become familiar with corpus linguistics methods during their studies.

**Future prospects**

Considering the development and use of conference interpreting corpora over the last three decades, further advances can be expected in terms of corpus features, availability, research, and applications.

Modern technology and collaborative efforts may lead to an increase in the number and size of conference interpreting corpora (Bendazzoli 2018). Some data sources, in particular, have already lent themselves to the establishment of large, machine-readable corpora, and these will continue to expand. In this respect, shared effort is required with a view to exchanges of methodological best practices, especially in data transcription and annotation.

Each approach to data collection entails both advantages and disadvantages for corpus compilers. While practisearchers may need to deal with issues relating to reflexivity, or expose themselves to possible conflicts with other interpreters reluctant to become the subject/object of academic research, external observations and indirect access inevitably offer more limited insight into the many variables involved. After all, source and target speeches in the form of transcripts remain a partial representation of spoken/signed communication and are the result of a selective process. This unavoidable limitation may be softened by greater inclusion of multimodal data in corpora, with direct linking to the relevant textual data. It is also hoped that technology may streamline the transcription, annotation, and alignment processes, so that even the less tech-savvy can readily exploit corpus methods. As regards annotation in particular, depending on the purposes behind the creation of a corpus, different types of annotation may be added along with extralinguistic details of the communicative situation, the role of participants and the speech events or text types involved. Whether a corpus is machine-readable or not, sharing (if not standardising) annotation systems would be a welcome improvement to allow for replications.

Corpus research may continue to pursue the objective of identifying interpreting norms and universals, though it is important to factor in the many variables (e.g. language specificity) that, even with large corpora, inevitably circumscribe any drawing of generalised conclusions. The description of typical and untypical patterns can nevertheless help stakeholders to have a better understanding of what it means to communicate through an interpreter, and can thus be a useful resource for informed debate on professional benchmarking and standards.

Conference interpreting education should continue to benefit from interpreting corpora. It would be interesting to promote the dissemination of pedagogical activities designed by individual researchers, with a view to fine-tuning relevant side-modules (e.g. targeted language enhancement, communication, linguistics, terminology) in support of interpreting classes *stricto sensu*. On the other hand, it is not clear when the impact of corpora on professional practice might start gaining ground—perhaps this may be seen more in the future, especially
Corpus studies in conference interpreting

among younger conference interpreters and in response to the evolution of working conditions that will almost certainly continue to gain pace in the years ahead.

Acknowledgements
I am grateful to Dr. Peter Gordon Mead for his invaluable suggestions while drafting this chapter.

Further reading

References
Claudio Bendazzoli


Bernardini, Silvia, Ferraresi, Adriano & Miličević, Maja 2016. From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective. Target 28 (1), 61–86.


Corpus studies in conference interpreting


Claudio Bendazzoli


