Chapter 1
Corpus-based Interpreting Studies: Past, Present and Future Developments of a (Wired) Cottage Industry

Claudio Bendazzoli

Abstract Drawing on Miriam Shlesinger's seminal paper on the creation and use of corpora in Interpreting Studies research, which she referred to as an *offshoot* of corpus-based translation studies (CTS) (Shlesinger 1998), and on Setton's chapter presenting an overview and prospects of Corpus-based Interpreting Studies (CIS) (Setton 2011), this chapter takes stock of nearly two decades of CIS and looks at the extent to which this offshoot has flourished in three areas of interpreting, namely research, education, and professional practice. Although considerable progress has been made in dealing with many of the challenges involved in corpus development, each area has reached a different degree of maturity with respect to CIS. Corpora are increasingly used in research, not only to look at simultaneous conference interpreting, but also to investigate other modes and settings, including consecutive interpreting and dialogue interpreting. This applies to both electronic, machine-readable corpora and more traditional datasets that are analyzed manually. On the other hand, their exploitation in interpreter education is still limited, while professional practice has yet to discover the potential of corpora. A better understanding of the reasons behind these discrepancies may be helpful to inform future directions of CIS and narrow the gap between research and (educational and professional) practice. Finally, the features of Web 2.0 are considered in order to discuss possible solutions to some of the methodological obstacles in the creation and use of interpreting corpora.

Keywords Corpus development · Corpus use · Manual analysis · Automatic analysis · Web 2.0

C. Bendazzoli (✉)
School of Management and Economics, Department of Economic and Social Studies, Mathematics and Statistics, University of Turin, Corso Unione Sovietica 218/bis, 10134 Torino, Italy
e-mail: claudio.bendazzoli@unito.it

M. Russo et al. (eds.), *Making Way in Corpus-based Interpreting Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-10-6199-8_1

1.1 Introduction

Computer technology has been having a dramatic 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 manually and detect all the occurrences of a certain phenomenon with the naked eye. From micro-analyses of short texts and small samples of talk, scholars now have the possibility to query millions of words making up different text types, be they written or spoken, that have been put together in a principled way to achieve a certain degree of representativeness (Biber et al. 1998; Renouf 2007).

As reported by Laviosa (2011, 2015, p. 31–36), such an unprecedented opportunity found fertile ground in (written) translation studies at the beginning of the 1990s (e.g. Baker 1993). It 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. Due to the intrinsic difficulties in gathering, transcribing and making spoken and sign-language data available in electronic form (Metzger and Roy 2011; Niemants 2012), the corpus-based approach began to be considered some time later in interpreting studies. Probably, the first paper about the idea of extending the corpus-based methodology to interpreting, as well as using already available monolingual corpora to test hypotheses about interpreting was published by Miriam Shlesinger in 1998. In this seminal work, Shlesinger refers to "corpus-based interpreting studies as an off-shoot of corpus-based translation studies" (ibid.), thus opening the way to this kind of research venture. More than ten years later, Robin Setton (2011) published a chapter which provides a broad overview of several corpus-based research projects, highlighting relevant methodological challenges and how these have been (or will have to be) tackled. And yet, in discussing the object and aims of this research paradigm, Setton states that "CIS is still a cottage industry" (2011, p. 34). Though the general sense of this idiomatic expression can be easily perceived even by non-native speakers of English, its precise meaning will be appreciated if we look it up in a dictionary. As reported in the Merriam-Webster online dictionary, a cottage industry is:

1. an industry whose labor force consists of family units or individuals working at home with their own equipment
2. a small and often informally organized industry
3. a limited but enthusiastically pursued activity or subject

The three definitions quoted above do mirror the status of corpus-based interpreting studies pretty well. Despite being more "limited" compared to CTS research, especially due to "small" corpus size, CIS as a research paradigm has been indeed pursued "enthusiastically" over the last two decades. The following section goes through Setton's overview and considers more recent initiatives accounting for an increasingly greater number of "family units or individuals" engaged in

corpus-based research. Then, Sects. 1.3 and 1.4 highlight how this field of inquiry may evolve from being informally organized to becoming more collaborative by taking advantage of the Internet as a platform to overcome some of its methodological obstacles and extend its applications to interpreter education and practice.

1.2 Nearly Twenty Years of Interpreting Corpora

Setton's overview of corpus-based interpreting research projects is interesting in many respects. Besides giving an account of the scholars, the interpreting modes, and the kind of enquiries involved, it offers a snapshot of how corpora have been built over time. At first glance, a general distinction into three broad categories can be made (Bendazzoli and Sandrelli 2009), i.e. manual corpora (not readily suitable for automatic extraction of occurrences), early machine-readable corpora, and fully machine-readable corpora (available to the scientific community). This distinction is here explored in greater detail to problematize the notion of corpus and how this has been intended in interpreting research. In addition to the works listed in Setton's overview, I consider further and more recent CIS projects, above all from Asian universities and research centers, as well as the projects presented at the first international CIS workshop held at the Department of Interpreting and Translation of the University of Bologna at Forlì in May 2015.[1] In the analysis below, I break down the main features that can be gleaned by looking at all these CIS projects.

1.2.1 Time Line

The first corpus-based research endeavor listed in Setton's overview is Oléron and Nanpon's work in 1965/2002, i.e. well before the seminal paper published by Miriam Shlesinger in 1998. At least seven more projects carried out before the new millennium are included in the same review, but it is unlikely that such early corpora were machine-readable, and probably corpus linguistic tools were not part of those studies. So why should they be classed as corpus-based? In fact, Setton focuses on "authentic corpora" (Setton 2011, p. 38) which means empirical data from real life interpreting assignments (not anecdotes, introspection, or experiments). On the other hand, among the subsequent corpora listed in the same overview together with the other projects taken into account in this chapter, an increasing number of projects are based on fully machine-readable transcripts (and in some cases these are tagged and indexed) and take advantage of automatic extraction of occurrences, thus they are

[1]The official name of the workshop is "Corpus-based Interpreting Studies: The State of the Art. First Forlì International Workshop". It was held on 7–8 May 2015 and gathered almost 100 participants. See http://eventi.dipintra.it/cis1/ for further details on the rationale of the event and the program.

fully in line with the classic definition of corpus 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, p. 9). Nevertheless, interpreting corpora created ad hoc by individual researchers for manual analysis are still present today and complement the realm of CIS. It is clear that the notion of corpus in Interpreting Studies has been initially linked to empirical research based on authentic data (i.e. from real life interpreting assignments) and that the difficulties in establishing electronic corpora have kept this notion applicable to data sets that continue to be analyzed manually. In fact, looking at various definitions of corpus proposed by linguists in the nineties, Tognini-Bonelli (2001, p. 52 ff.) foregrounds authenticity, representativeness, and sampling criteria as main points, and yet she too gives a definition making reference to a "computerized collection of authentic texts, amenable to automatic or semi-automatic processing or analysis" (ibid. p. 55).

1.2.2 Languages

The CIS projects under consideration cover a wide range of language combinations and confirm one of the "special challenges" (Setton 2011, p. 68) of CIS, i.e. multilingualism (the other challenges are orality, situatedness and immediacy). As could be expected, English is represented in many studies, but it is encouraging to see that corpus linguistics tools are available to annotate and analyze non-European languages too (e.g. Hebrew, Chinese, Japanese).

1.2.3 Interpreting Modes and Settings

All the pioneering works are focused on professional simultaneous interpreting performed in conference settings. This trend remains constant throughout the years, and becomes particularly strong in relation to two specific sources of data, i.e. TV broadcasting and the European Parliament (EP). The latter, in particular, has granted the opportunity to develop major CIS projects (e.g. EPIC in Monti et al. 2005). Due to the abundant availability of source and target speeches translated simultaneously into as many as 23 different languages, and the permission to use them for research purposes, this setting has lent itself to multiple enquiries and will continue to represent a valuable data source (Bendazzoli 2010). Despite the appeal of EP data, further conference settings have been explored, e.g. festivals, medical conferences, football press conferences and many others. On the other hand, among the studies from Asian research centers there is a prevalence of consecutive interpreting (CI) over simultaneous interpreting (SI) corpora. The reason is linked, once again, to data accessibility. In fact, all the research projects on CI are based on the same data source, i.e. televised press conferences of Chinese political representatives (Wang 2015). In addition, more recent projects also concern short consecutive

interpreting in community settings or dialogue interpreting, and promising efforts are also being made to develop sign language interpreting corpora, though there are greater obstacles to collect this kind of data due to confidentiality and anonymity[2] issues (see Metzger and Roy 2011 for sign language interpreting).

1.2.4 Corpus Size

Interpreting corpora are still quite limited in terms of size if compared to general reference corpora such as the spoken part of the British National Corpus (Dembry and Love 2015). Even the projects based on EP data have yet to reach the size of general reference corpora, though it may well be just a matter of time and labor force as there is potential to arrive at sufficiently large resources (i.e. including millions of tokens) to make generalizations about simultaneous interpreting in the EP context. Despite the advantages of data collection from this source, it is hoped that other international organizations become the object of CIS research in order to diversify the types of setting and improve the generalizability of the results. Reassuringly, there are studies that consider other international organizations, such as the European Commission (Spinolo this volume; Scardulla 2016) and the United Nations (Dayter 2016).

According to the literature on Interpreting (Matsubara et al. 2002; Ono et al. 2008; Tohyama et al. 2005), the Simultaneous Interpretation Database (SIDB, or CIAIR as reported in Setton's overview) of Nagoya University is the only interpreting corpus containing one million words to date, though it has not been possible to obtain further details and access this large interpreting corpus.

Ongoing works, especially in Asia, are expected to deliver sizeable resources, but insofar as CIS projects are the responsibility of individual researchers, it is unlikely that very large corpora can become available in the near future.[3]

1.2.5 Transcription, Data Annotation and Analysis

As mentioned above, all the early CIS projects involve manual data analysis, which continues to be performed at present but with some form of (manual) annotation to

[2]Different strategies can now be adopted to anonymize video data, e.g. blurring, pixelating, adding a bar mask over the eyes, altering the color settings to obtain a negative picture visualization, masking identity through avatars that replicate the same facial expressions, but these processes are demanding in terms of time and resources, and they may also alter the data considerably (see Corti et al. 2014, p. 122–124; Gellerman 2016). Licoppe and Veyrier (2016) simply use the "edge detection effect" available in Movie Maker to alter color patterns and anonymize data from remote court interpreting videos (Veyrier, personal communication).

[3]This is not to undervalue the role of small and specialized corpora (see Ghadessy et al. 2001).

categorize the phenomena under investigation. Machine-readable corpora instead rely on automatic or semi-automatic annotation as transcripts are tagged by means of different software programs, e.g. Treetagger and CLAWS for Part-of-Speech tagging (see Sandrelli and Bendazzoli 2006; Dayter 2016) or Exmarlda (Schmidt and Wörner 2009), and apply conventions that tend to be shared to a larger extent, e.g. TEI (as in Cencini 2002; Cencini and Aston 2002) or CHAT/CLAN (as in Álvarez de la Fuente and Fernández Fuertes 2014).

Generally, there is a close relationship between the type of transcription system adopted and the kind of annotation applied. This calls not so much for a uniform standard, which is utopian in many respects,[4] but for a basic transcription format in which the verbal level is represented without too many details from the paralinguistic and kinetic levels. This way it should be possible to import transcripts onto different tools and apply encoding systems without excessive adjustment work, while further levels of annotation can be added at a later stage. After all, "[e]very corpus will have limitations, but a well-designed corpus will still be useful for investigating a variety of linguistic issues" (Biber et al. 1998, p. 250). Keeping such basic transcription format as a basis for future annotations I believe is part of effective corpus design to assure greater exploitation by multiple researchers.

1.2.6 Data Availability and Distribution

Starting with the oldest works listed in Setton's overview, it can be noticed that transcripts are hardly available now and sound files were recorded on tape, not in digital form. Limited access to transcript files is also reported in subsequent works, such as Diriker (2004) in which the analysis is still carried out manually. On the other hand, Cencini (2002) and Fumagalli (2000) appear as the first cases in which computer-aided inquiries are performed on machine-readable transcripts: the former looks at TV interpreting, applying a TEI-conformant encoding system; the latter investigates consecutive interpreting by means of concordance software Multiconcord (see Corness 2002), and yet the transcripts are "not available for outside use" (Setton 2011, p. 40). However, the data used in the majority of the subsequent studies reported by Setton can be made available to other researchers, either on CDs (e.g. Vuorikoski 2004; Monacelli 2009) or on the web, as is the case with the corpus projects illustrated in Angermeyer et al. (2012), with EPIC[5] (Russo et al. 2012), and with DIRSI-C (Bendazzoli 2012). Given the many efforts needed to set up an interpreting corpus, the last step in corpus development, i.e. access and

[4]This applies not only to CIS but more generally to spoken corpora (Schmidt 2014).

[5]Only the indexed, POS-tagged, and lemmatized transcripts of this corpus can be queried online, but the entire corpus, including the text files of individual transcripts, the video files of source speeches and the audio files of target speeches can be obtained from the European Language Resources Association (ELRA).

distribution, should not be underestimated and researchers should strive to obtain permission to share the data at least within the research community.

1.2.7 General Remarks

At the end of this extended review of corpus-based interpreting research projects, it can be concluded that in the past CIS scholars could not count on ready-made corpora and largely had a DIY (i.e. do-it-yourself) approach to developing corpora (McEnery et al. 2006, p. 71). Moreover, considering the degree of access to recordings and transcripts, this approach can be seen as a do-it-yourself and *keep-it-for-yourself* experience with very limited use of computational linguistics tools. Looking at more recent studies, it can be argued that CIS scholars have continued to build their own language resources in a do-it-yourself fashion, but not just for themselves or by themselves. In this second stage, analyses are carried out both manually and by means of corpus linguistics tools – in fact the two approaches are not to be considered mutually exclusive in that supplementing quantitative analyses with qualitative explorations can lead to even more robust results. Furthermore, the exclusive focus on simultaneous (conference) interpreting has been expanded to touch upon other modes and settings, including classic consecutive and short consecutive (dialogue) interpreting. In addition to the opportunities afforded by technological progress in data collection and data processing, it seems that greater awareness of interpreting research as well as collaboration between *practisearchers* and the relevant communities of practice have favored these positive developments. The challenge now lies in feeding such communities of practice with the results obtained from research so as to increase their awareness, improve standards and find possible applications to interpreter education and professional practice.

1.3 More Than Research?

One of the advantages of corpus-based research is that it is possible to analyze large data sets and extract relevant occurrences automatically, thus looking at trends, patterns, concordances, common and uncommon phenomena emerging from a representative sample of language use. When it comes to Translation in general, and interpreting in particular, comparisons can be made between source and target texts/speeches (parallel analysis) or between texts and speeches in the same language but produced in different ways (comparable analysis), and take into account specific language pairs from a unidirectional or bidirectional perspective (Laviosa 2012). Interpreters' output can also be investigated as a kind of discourse in its own right to point out features that differentiate it from other types of discourse (e.g. with speakers operating under different conditions and in different settings). These and

many more lines of research can provide highly interesting insight into both the process and the product of language mediation activities.

In addition, corpora can be used for educational purposes, and there is now an increasingly established tradition in language learning (see, for instance, Aston 2001; Braun 2006; Flowerdew 2012; Gavioli 2005; Kohn 2012) with tangible effects on curriculum design (McCarthy and McCarten 2012; Prat Zagrebelsky 2004). Corpus applications have also found fertile ground in translator training and education (Beeby et al. 2009; Bowker 2003; Zanettin et al. 2003) although more awareness among educators and students would be needed, and more user-friendly tools for corpus building and interrogation are required to have a stronger impact and reach professional communities as well (Bernardini and Castagnoli 2008). As regards interpreter training and education, the potential use of corpus-based and corpus-driven methods is just in its infancy. Most of the projects taken into account in this chapter are designed for research purposes. However, some training and education-related initiatives can be accounted for. Firstly, the availability of ready-made interpreting corpora is a convenient way for students to obtain data for their graduation theses and carry out small-scale investigations leading to greater self-awareness of some of the processes and phenomena they have encountered in their curriculum (Dal Fovo 2011; Russo 2010). Secondly, corpus concordances can offer relevant materials to increase trainees' awareness of certain phraseological units and other language-specific features that may be difficult to acquire and automatize (see Aston this volume, 2016; Bale 2013; Lázaro Gutiérrez and del Mar Sánchez Ramos 2015; Sandrelli 2010). Thirdly, learner corpora involving interpreting students (see, for instance, Leung and Yip 2013; Niemants 2013) are also valuable resources not only to ascertain the kinds of difficulties and challenges trainees face during the acquisition and development of mediation skills, but also to promote self-assessment among learners and set up repositories of pedagogical materials. Indeed, interpreting corpora are based on multimedia archives, which generally include videos or audio recordings plus transcripts from different domains. Trainees can take advantage of such repositories to be exposed to a variety of source and target speeches and for extra practice. Some of these corpora are designed with a specific educational challenge in mind, e.g. source speakers using English as a lingua franca (Pignataro 2014), or they gather a wider range of materials which are nonetheless classified on the basis of a number of parameters. For example, the speech repository designed by the DG Interpretation of the European Commission[6] can be searched on the basis of seven parameters. These are *language* (29 languages are available), *level* (from basic to beginner, intermediate, advanced/test-type, and very advanced), *use* (simultaneous or consecutive), *domain* (52 topics concerning EU policy areas plus a general category for other topics), *type* (six different communicative situations such as conference, press conference, debate, hearing, interview, and pedagogical material), *keyword* and *speech number*. Not all the speeches are supplemented with their transcript, so it might be

[6]See https://webgate.ec.europa.eu/sr/ (accessed 7 October 2016).

questionable to consider this repository a corpus according to the definition presented above. However, all these classification parameters do serve as annotations of features that make these materials distinguishable from others and may well function as metadata normally accounted for when compiling a corpus. The search parameters listed above are described on the speech repository website and were designed by experienced interpreters and trainers. However, other methods of classification are also possible, for example allowing users to assign certain tags to a speech on the basis of the perceived difficulty, following a bottom-up approach that is typical of Web 2.0 applications. In the next section I present the main features of Web 2.0 and argue that some of the obstacles to creating interpreting corpora may be overcome thanks to the advantages offered by this new version of the web and its related applications.

As regards professional practice, the question whether corpora can play a role is still open and remains to be addressed. I cannot provide survey data in this respect, but if corpora are still largely unknown by translators (Bernardini and Castagnoli 2008) this is likely to be the case, even more so, amongst interpreters. Even those who might have been exposed to corpus linguistics during their training and education would hardly consider consulting corpora or creating ad hoc ones when preparing for or while working in an assignment due to the time constraints typical of interpreter-mediated settings. However, interpreters may be familiar with Translation Memories, terminological databases, and glossaries. If they had user-friendly tools at their disposal that are capable of processing textual data (e.g. word documents, power point slides, pdf files, websites and so on) to quickly build corpora from which word lists, keywords, collocations, and n-grams, just to mention some examples, could be obtained, perhaps more interest would be raised in this community of practice. Bootstrapping techniques (Baroni and Bernardini 2004) have already been introduced in translator training and education to create ad hoc corpora from the web allowing translators to browse these resources, review terminology and retrieve background information, especially when translating into a foreign language. The same technique has also been proposed for interpreters' preparation prior to assignments, highlighting the time-saving advantage over manual terminological extraction (Fantinuoli 2006). More user-friendly tools are also being developed, e.g. text analysis software TranslatorBank and, more specifically, InterpretBank[7] (Fantinuoli 2012). These tools are more intuitive than traditional corpora. They enable users to create specialized corpora quickly from the web and any assignment-related materials delivered by the client in advance, thus making it easier and faster to create glossaries and manage terminological databanks on the fly.

[7]See http://www.staff.uni-mainz.de/fantinuo/translatorbank.html and http://www.interpretbank.com/ (accessed 23 October 2016).

1.4 The Role of Web 2.0

The notion of Web 2.0 was put forward at the beginning of the new millennium when IT experts realized that a new version of the Web had become available, e.g. in the form of sites and services such as Wikipedia, Facebook, MySpace, Delicious, You Tube, Fickr, blogs, Google and the like (McAfee 2009). The main difference from the previous version of the Web is that the Internet started to be used as a platform: data sharing, user collaboration and interaction, content creation and exchange have become possible to the point that information is not only accessed, but also actively and collaboratively created thus generating emergent patterns of use without any pre-imposed structure. The new technologies and applications involved have the power of "bring[ing] people together and let them interact, without specifying how they should do so" (McAfee 2009, p. 2). For example, this is clear in the way web users apply tags to content that has been posted online: whether it is just a "like" or a more sophisticated rating system, the emerging (bottom-up) classification is identified as a folksonomy, i.e. "a categorization system developed over time by folks. A folksonomy is an alternative to a taxonomy, which is a categorization system developed at a single point in time by an authority" (ibid., p. 73).

The pervasiveness of this approach is evident in social media, though McAfee foregrounds the term "collaborative" when applying Web 2.0 to business in what he calls Enterprise 2.0 (ibid., p. 16). Here I propose to consider a similar application to Interpreting Research and, in particular, to CIS, as the advantages of Web 2.0 could be helpful to face some of the challenges typically present in developing interpreting corpora, e.g. in data collection, transcription, annotation, and distribution.

I would like to argue that CIS can benefit indeed from the potential provided by Web 2.0 at different levels:

- in data collection, Web 2.0 could be exploited to give greater visibility to research projects, and their objectives should be beneficial in some way to the communities from which data are taken;
- in corpus development, transcribing could be done collaboratively by multiple teams;
- in annotation, certain types of tags could be added subsequently by other researchers or by a target group of corpus/interpreting service users;
- in distribution, corpora could be exchanged (and used for replications or other investigations) more easily, thus fostering "the development of corpus construction tools and dissemination platforms that can enable researchers to archive their resources and make them available to others" (Ruhi et al. 2014, p. 8) which is considered a "pressing concern […] in spoken corpora research" (ibid.).

Web 2.0 applications for managing media resources that are also used in interpreter training (e.g. speech repositories) have been developed over the last few years. For example, the University of Geneva created SIMON (Sharing Interpreting

Materials Online), which is a platform for interpreter trainers to exchange and create pedagogical materials (Seeber 2006). Another example is Speechpool.[8] This online repository allows the user to retrieve speeches on the basis of how popular they are amongst other users and how recently the speeches have been uploaded in a bottom-up approach. Other parameters are similar to the ones found in the EU speech repository, e.g. *topic* (37 options plus a general option for uncategorized speeches) and *suitable for*, which proposes different modes (consecutive without notes, consecutive, advanced consecutive, simultaneous, and advanced simultaneous). This is an example of Web 2.0 resource, as web users themselves contribute to content creation and sharing, and to the classification of the materials letting patterns emerge from the choices they make when selecting the speeches. While this form of independent learning takes advantage of corpora, it is still far from providing a clear integration of CIS resources into an interpreting curriculum. The few examples mentioned above are nonetheless an encouraging starting point and hopefully there will be more in the near future.

1.5 Conclusion

In this chapter I took Shlesinger's seminal paper on corpus-based interpreting studies (1998) and Setton's comprehensive overview of corpus-based interpreting research projects (2011) as a starting point to offer an update on the development and applications of interpreting corpora over the last two decades. General developmental trends have been pointed out showing that the corpus-based approach has been increasingly used across different interpreting modes and settings: from simultaneous interpreting at conferences to international institutions such as the European Parliament, from televised consecutive interpreting assignments during governmental press conferences in China to face to face or over the phone community interpreting. The degree of data accessibility and confidentiality obviously have a strong bearing on the amount of data that can be accessed in each setting as well as on the effort required of the researcher to process them to make up a representative corpus. Indeed, the term *corpus* continues to be applied to principled collections of data regardless of their machine-readability, especially in case of DIY corpora created by individual scholars (still quite common in CIS).

In addition to a variety of research efforts, corpus-based and corpus-driven applications are slowly finding their way in interpreter training/education and practice. Initial examples of corpus use in the interpreting curriculum offer stimulating ways to increase trainees' awareness of collocations, phraseology, language specific features, and terminological preparation. The latter is now supported by corpus applications through tools that are being made more intuitive and

[8]See http://www.speechpool.net/en (accessed 23 October 2016).

user-friendly in order to quickly create corpora from the web or background documentation (e.g. slides, reports, etc.).

In conclusion, the way forward in CIS is likely to see not only more corpora, some of them much larger in size than is currently the case, but also corpora on further types of interpreting as a result of greater collaboration in data collection, transcription, annotation, and sharing. Interpreting corpora, whether machine-readable or not, are valuable language resources as they do not serve just the purpose of an individual research project, they also have enormous potential in the development of interpreter training/education and professional practice. As pointed out by Fantinuoli and Zanettin (2015, p. 8):

> Corpus-based translation studies have steadily grown as a disciplinary sub-category since the first studies began to appear more than twenty years ago. A bibliometric analysis of data extracted from the Translation Studies Abstracts Online database shows that in the last ten years or so about 1 out of 10 publications in the field has been concerned with or informed by corpus linguistics methods (Zanettin et al. 2015).

A similar trend for the future can be expected of corpus-based interpreting studies, even if manual analysis and small corpus size will continue to be fundamental features of this cottage industry – a cottage industry that is nevertheless turning increasingly wired, as it takes advantage of the potential of Web 2.0 technologies and collaborative work leading to larger and more representative interpreting corpora.

Appendix

The following table lists the CIS projects considered in the present chapter in addition to the works included in the overview by Setton (2011) and those presented at the conference *Corpus-based Interpreting Studies: The State of the Art. First Forlì International Workshop*, which was held at the University of Bologna (Forlì campus) on 7–8 May 2015 (see http://eventi.dipintra.it/cis1/):

Language Resource/Reference	Languags	Setting/Communicative situation	Interpreting mode	Subjects	Length	Transcription published or available	Sound files availability	Analysis
Leung and Yip (2013)	EN > < ZH	Interpreter training classes		9 trainees		Online web interface http://arts.hkbu.edu.hk/~engester/main.html		
Taehyung (2011)	EN > KO	Academy awards ceremony (TV)	SI and live captions					TV viewers' preference for SI *vs.* live captions
CEIPPC (Wang 2012a)	ZH > EN	Premier press conferences (TV)	CI	5 pros (5 conferences)	71,730 words	Video rec. not yet available to other scholars		Shifts in TT (addition, reduction, correction)
CEIPPC (Wang 2012b)	ZH > EN	Premier press conferences (TV)	CI	7 pros (8 conferences)	Over 100,000 words	ParaConc, same as above		Same as above
Court interpreting corpus (Biagini 2012)	FR <>IT	Italian courtrooms (Turin and Pisa)	CI	6 interpreters	7 hearings (approx.. 9 h)	Manual analysis (CA approach)		
CECIC (Hu and Tao 2013; Hu 2016, p. 197–221)	ZH > EN ZH > EN EN	-Premier press conferences (TV) -Government's written reports (web) -CNN press conferences (web)	CI Translation (org-en)	Parallel and Comparable corpus	544,211 (int TT 96,205 from 133,431 ST)	POS-tagged (Treetagger and ICTCLAS 3.0)		Features of EN TT, hedging (*some*), delexical verbs (*make*)
Raquel Lázaro Gutiérrez, María del Mar Sánchez Ramos (2015)	ES	Written texts + video transcripts (simulated and real conversations) on gender violence	pedagogical tool for public service interpreting training			xml tagging; to be used with different tools		Pragmatic content in gender violence genre. Pedagogical purposes

(continued)

(continued)

Language Resource/Reference	Languags	Setting/Communicative situation	Interpreting mode	Subjects	Length	Transcription published or available	Sound files availability	Analysis
Consecutive interpreting notes corpus (Kellet Bidoli 2016; Vardè 2014)	IT > EN EN > IT	Experimental	CI	5 beginners + 5 advanced students + 5 trainers	60 target speeches + notes	Livescribe smartpen		Problems and strategies in reception stage
Fu (2016)	ZH > EN	Chinese premier press conferences + reports 2008–2012	CI/SI Translation	pros	2 h	WordSmith		Modality
SINC, student interpreter narrative corpus (Voinova and Ordan 2016)	HE	Weekly reports + end of year assignments	NA	73 student interpreters	288,000 words	Sketch Engine compatible format		Narratives by students attending community interpreting course

References

Álvarez de la Fuente, Esther, and Raquel Fernández Fuertes. 2014. A methodological approach to the new analysis of natural interpreting: Bilingual acquisition data and the CHAT/CLAN tool/Un enfoque metodológico para el análisis de la interpretación natural: los datos de adquisición bilingüe y la herramienta CHAT/CLAN". In *Corpus-based Translation and Interpreting Studies: From description to application/Estudios traductológicos basados en corpus: de la descripción a la aplicación*, ed. M.T. Sánchez Nieto, 77–104. Berlin: Frank & Timme.

Angermeyer, Philipp Sebastian, Bernd Meyer, and Thomas Schmidt. 2012. Sharing community interpreting corpora: A pilot study. In *Multilingual corpora and multilingual corpus analysis*, ed. T. Schmidt, and K. Wörner, 275–294. Amsterdam: John Benjamins.

Aston, Guy. 2016. How corpora can help the interpreter walk the tightrope. In *Corpus-based approaches to translation and interpreting: From theory to applications*, ed. G. Corpas Pastor, and M. Seghiri. Frankfurt: Peter Lang.

Aston, Guy (ed.). 2001. *Learning with corpora*. Bologna: CLUEB.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and technology: In honour of John Sinclair*, ed. M. Baker, G. Francis, and E. Tognini-Bonelli, 233–250. Amsterdam: John Benjamins.

Bale, Richard. 2013. Undergraduate consecutive interpreting and lexical knowledge: The role of spoken corpora. *The Interpreter and Translator Trainer* 7 (1): 27–50.

Baroni, Marco, and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of LREC 2004, fourth international conference on language resources and evaluation. Lisbon – Portugal 26–28 May 2004*, eds. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva, with the collaboration of C. Pereira, F. Carvalho, M. Lopes, M. Catarino, and S. Barros, 1313–1316. Paris: ELRA/ELDA.

Beeby, Allison, Rodríguez Inés Patricia, and Pilar Sánchez-Gijón (eds.). 2009. *Corpus use and translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam: John Benjamins.

Bendazzoli, Claudio, and Annalisa Sandrelli. 2009. Corpus-based interpreting studies: Early work and future prospects. *Revista Tradumàtica. L'aplicació dels corpus lingüístics a la traducció* 7. https://dialnet.unirioja.es/servlet/articulo?codigo=5098399. Accessed 24 June 2016.

Bendazzoli, Claudio. 2010. The European Parliament as a source of material for research into simultaneous interpreting: Advantages and limitations. In *Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungenzur Translationswissenschaft VI (Forum Translationswissenschaft, Bd. 12)*, ed. N.L. Zybatow, 51–68. Frankfurt a. M.: Peter Lang.

Bendazzoli, Claudio. 2012. From international conferences to machine-readable corpora and back: An ethnographic approach to simultaneous interpreter-mediated communicative events. In *Breaking ground in corpus-based Interpreting Studies*, ed. F. Straniero Sergio, and C. Falbo, 91–117. Frankfurt a. M.: Peter Lang.

Bernardini, Silvia, and Sara Castagnoli. 2008. Corpora for translator education and translation practice. In *Topics in language resources for translation and localisation*, ed. E. Yuste, 39–55. Amsterdam: John Benjamins.

Biagini, Marta. 2012. Data collection in the courtroom: Challenges and perspectives for the researcher. In *Breaking ground in corpus-based interpreting studies*, ed. F. Straniero Sergio, and C. Falbo, 231–251. Frankfurt a. M.: Peter Lang.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bowker, Lynne. 2003. Corpus-based applications for translator training: Exploring the possibilities. In *Corpus-based approaches to contrastive linguistics and translation studies*, ed. S. Granger, J. Lerot, and S. Petch-Tyson, 169–183. Amsterdam: Rodopi.

Bowker, Lynne, and Jennifer Pearson. 2002. *Working with specialized language. A practical guide to using corpora*. London: Routledge.

Braun, Sabine. 2006. ELISA–a pedagogically enriched corpus for language learning purposes. In *Corpus technology and language pedagogy: New resources, new tools, new methods*, ed. S. Braun, K. Kohn, and J. Mukherjee, 25–47. Frankfurt a. M.: Peter Lang.

Cencini, Marco, and Guy Aston. 2002. Resurrecting the corp(us/se): Towards an encoding standard for interpreting data. In *Interpreting in the 21st century. Challenges and opportunities. Selected papers from the first Forlì conference on Interpreting Studies, 9–11 November 2000*, eds. G. Garzone, and M. Viezzi, 47–62. Amsterdam: John Benjamins.

Cencini, Marco. 2002. On the importance of an encoding standard for corpus-based interpreting studies: Extending the TEI scheme. *CULT2 K. Special Issue of InTRALinea*. http://www.intralinea.org/specials/article/1678. Accessed 26 November 2012.

Corness, Patrick. 2002. Multiconcord. A computer tool for cross-linguistic research. In *Lexis in contrast. Corpus-based approaches*, ed. B. Altenberg, and S. Granger, 307–326. Amsterdam: John Benjamins.

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woolard (eds.). 2014. *Managing and sharing research data. A guide to good practice*. London: Sage.

Dal Fovo, Eugenia. 2011. Through the CorIT looking-glass and what MA students found there. *The Interpreters' Newsletter* 16: 1–20.

Dayter, Daria. 2016. Corpus-based approach to simultaneous interpretation at the United Nations: Multidimensional analysis of variation. Paper presented at the EST Congress 2016, Aarhus, Denmark, 15–17 September 2016.

Dembry, Claire, and Robbie Love. 2015. *Collecting the Spoken BNC2014 – overview of methodology*. Paper presented at the Corpus Linguistics 2015 Conference, Lancaster University, UK, 21–24 July 2015.

Diriker, Ebru. 2004. *De-/Re-contextualizing conference interpreting: Interpreters in the ivory tower?*. Amsterdam: John Benjamins.

Fantinuoli, Claudio. 2006. Specialized corpora from the Web and term extraction for simultaneous interpreters. In *Wacky! Working papers on the Web as Corpus*, ed. M. Baroni, and S. Bernardini, 173–190. Bologna: GEDIT.

Fantinuoli, Claudio. 2012. *InterpretBank – Design and implementation of a terminology and knowledge management software for conference interpreters*. PhD Thesis, Johannes Gutenberg University Mainz, Germersheim.

Fantinuoli, Claudio, and Federico Zanettin (eds.). 2015. *New directions in corpus-based translation studies*. Berlin: Language Science Press.

Flowerdew, Lynne. 2012. Corpora in the classroom: An applied linguistic perspective. In *Corpus applications in applied linguistics*, ed. K. Hyland, C. MengHuat, and M. Handford, 208–224. London: Bloomsbury.

Fu, Rongbo. 2016. Comparing modal patterns in Chinese-English interpreted and translated discourses in diplomatic setting. A systemic functional approach. *Babel* 62 (1): 104–121.

Fumagalli, Daniela. 2000. *Alla ricerca dell'interpretese. Uno studio sull'interpretazione consecutiva attraverso la corpus linguistics*. Unpublished MA Thesis, Advanced School for Translators and Interpreters (SSLMIT), University of Trieste.

Gavioli, Laura. 2005. *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

Gellerman, Helena. 2016. *What are the main issues with anonymization and feature extraction?* Paper presented at the FOT-NET Data Workshop on Data Anonymization and Feature Extraction, August 31-September 1, SAFER Vehicle and Traffic Safety Centre, Gothenburg.

Ghadessy, Mohsen, Alex Henry, and Robert L. Roseberry (eds.). 2001. *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins.

Hu, Kaibao. 2016. *Introducing corpus-based translation studies*. New York: Springer.

Hu, Kaibao, and Qing Tao. 2013. The Chinese-English conference interpreting corpus: Uses and limitations. *Meta* 58 (3): 626–642.

Kellett Bidoli, J. Cynthia. 2016. Methodological challenges in Consecutive Interpreting Research: Corpus analysis of notes. In *Addressing methodological challenges in Interpreting Studies Research*, eds. C. Bendazzoli, and C. Monacelli, 141–169. Newcastle upon Tyne: Cambridge Scholars Publishing.

Kohn, Kurt. 2012. Pedagogic corpora for content and language integrated learning. Insights from the BACKBONE Project. *The Eurocall Review* 20 (2): 3–22.

Laviosa, Sara. 2011. Corpus-based translation studies: Where does it come from? Where is it going? In *Corpus-based Translation Studies. Research and applications*, eds. A. Kruger, K. Wallmach, and J. Munday, 13–32. London: Continuum.

Laviosa, Sara. 2012. Corpora and translation studies. In *Corpus applications in Applied Linguistics*, eds. K. Hyland, C. Meng Huat, and M. Handford, 67–83. London: Bloomsbury.

Laviosa, Sara. 2015. Corpora and holistic cultural translation. In *Corpus-based Translation and Interpreting Studies: From description to application/Estudios traductológicos basados en corpus: de la descripción a la aplicación*, ed. M.T. Sánchez Nieto, 31–51. Berlin: Frank & Timme.

Lázaro Gutiérrez, Raquel, and María del Mar Sánchez Ramos. 2015. Corpus-based interpreting studies and public service interpreting and translation training programs: The case of interpreters working in gender violence contexts. In *Yearbook of Corpus Linguistics and Pragmatics 2015. Current approaches to discourse and Translation Studies*, ed. J. Romero-Trillo, 275–292. Cham: Springer.

Leung, S.M. Ester, and Leonard Yip. 2013. *A bilingual corpus of interpreting students' performance*. http://arts.hkbu.edu.hk/∼engester/main.html . Accessed 21 Sept 2016.

Licoppe, Christian, and Clair-Antoine Veyrier. 2016. Consecutive courtroom interpreting and the management of long turns: Video-mediated hearings at the French appeal court for asylum demands. Paper presented at the 6th International Conference on Applied Linguistics and Professional Practice (ALAPP) "Transnational flows and professional practice", University of Copenhagen, Denmark, 3–5 November 2016.

Matsubara, Shigeki, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2002. Bilingual spoken monologue corpus for simultaneous machine interpretation research. In *LREC 2002. Proceedings of the third international conference on language resources and evaluation, 29th, 30th &31st May 2002, Las Palmas de Gran Canaria. Vol. 1*, eds. M. González Rodríguez, and C.P. Suárez Araujo, 153–159. Paris: ELRA.

McAfee, Andrew. 2009. *Enterprise 2.0. New collaborative tools for your organization's toughest challenges*. Boston: Harvard Business Press.

McEnery, Tony, Richard Xiao, and Tono Yukio. 2006. *Corpus-based languages. An advanced resource book*. London: Routledge.

Metzger, Melanie, and Cynthia Roy. 2011. The first three years of a three-year grant. When a research plan doesn't go as planned. In *Advances in interpreting research: Inquiry in action*, ed. B. Nicodemus, and L. Swabey, 59–84. Amsterdam: John Benjamins.

McCarthy, Michael, and Jeanne McCarten. 2012. Corpora and materials design. In *Corpus applications in applied linguistics*, ed. K. Hyland, C. Meng Huat, and M. Handford, 223–241. London: Bloomsbury.

Monacelli, Claudia. 2009. *Self-preservation in simultaneous interpreting. Surviving the role*. Amsterdam: John Benjamins.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli, and Mariachiara Russo. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta* 50 (4). http://id.erudit.org/iderudit/019850ar . Accessed 4 May 2016.

Niemants, S.A. Natacha. 2012. The transcription of interpreting data. *Interpreting* 14 (2): 165–191.

Niemants, S.A. Natacha. 2013. From role-playing to role-taking: Interpreter role(s) in healthcare. In *Interpreting in a changing landscape: Selected papers from Critical Link*, ed. C. Schaeffner, K. Kredens, and Y. Fowler, 305–319. Amsterdam: John Benjamins.

Oléron, Pierre, and Hubert Nanpon. 1965/2002. Research into simultaneous translation. In *The interpreting studies reader*, ed. F. Pöchhacker, and M. Shlesinger, 43–50. London: Routledge.

Ono, Takahiro, Hitomi Tohyama, and Matsubara Shigeki. 2008. Construction and analysis of word-level time-aligned simultaneous interpretation corpus. In *Proceedings of the sixth international conference on language resources and evaluation (LREC '08)*, eds. N. Calzolari,

K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/. Accessed 13 Jan 2014.

Pignataro, Clara. 2014. ELF pragmatics and interpreting/Pragmática del inglés como lengua franca e interpretación. In *Corpus-based Translation and Interpreting Studies: From description to application/Estudios traductológicos basados en corpus: de la descripción a la aplicación*, ed. M.T. Sánchez Nieto, 105–124. Berlin: Frank.

Renouf, Antoinette. 2007. Corpus development 25 years on: From super-corpus to cyber-corpus. In *Corpus Linguistics 25 years on*, ed. R. Facchinetti, 27–49. Amsterdam: Rodopi.

Ruhi, Şükriye, Michael Haugh, Thomas Schmidt, and Kai Wörner (eds.). 2014. *Best practices for spoken corpora in linguistic research*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Russo, Mariachiara. 2010. Reflecting on interpreting practice: Graduation theses based on the European Parliament Interpreting Corpus (EPIC). In *Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungenzur Translationswissenschaft VI (Forum Translationswissenschaft, Bd. 12)*, ed. L.N. Zybatow, 35–50. Frankfurt a. M.: Peter Lang.

Russo, Mariachiara, Claudio Bendazzoli, Annalisa Sandrelli, and Nicoletta Spinolo. 2012. The European Parliament Interpreting Corpus (EPIC): Implementation and developments. In *Breaking ground in corpus-based Interpreting Studies*, eds. F. Straniero Sergio, and C. Falbo, 53–90. Frankfurt a. M.: Peter Lang.

Sandrelli, Annalisa. 2010. Corpus-based Interpreting Studies and interpreter training: A modest proposal. In *Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungenzur Translationswissenschaft VI (Forum Translationswissenschaft, Bd. 12)*, ed. L.N. Zybatow, 69–90. Frankfurt: Peter Lang.

Sandrelli, Annalisa, and Claudio Bendazzoli. 2006. Tagging a corpus of interpreted speeches: The European Parliament Interpreting Corpus (EPIC). In *Proceedings of the LREC 2006 Conference, Genova, Magazzini del Cotone 24–26 May 2006*. Genova: ELRA.

Scardulla, Cristina. 2016. ELF interpreting at the European Union: A corpus-based study. Paper presented at the EST Congress 2016, Aarhus, Denmark, 15–17 September 2016.

Schmidt, Thomas. 2014. (More) common ground for processing spoken language corpora? In *Best practices for spoken corpora in linguistic research*, ed. S. Ruhi, M. Haugh, T. Schmidt, and K. Wörner, 249–265. New Castle upon Tyne: Cambridge Scholars Publishing.

Schmidt, Thomas, and Kai Wörner. 2009. EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19 (4): 565–582.

Seeber, G. Kilian. 2006. SIMON: An online clearing house for interpreter training materials. In *Proceedings of Society for Information Technology and teacher education international conference*, ed. C. Crawford, R. Carlsen, K. McFerrin, J. Price, R. Weber, and D.A. Willis, 2403–2408. Chesapeake: AACE.

Setton, Robin. 2011. Corpus-based interpreting studies (CIS): Overview and prospects. In *Corpus-based translation studies. Research and applications*, ed. A. Kruger, K. Wallmach, and J. Munday, 33–75. London: Continuum.

Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43 (4): 486–493.

Taehyung, Lee. 2011. English into Korean simultaneous interpretation of Academy Awards Ceremony through open captions on TV. *Meta* 56 (1): 145–161.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tohyama, Hitomi, Shigeki Matsubara, Nobuo Kawaguchi, and Inagaki Yasuyoshi. 2005. Construction and utilization of bilingual speech corpus for simultaneous machine interpretation research. In *Proceedings of 9th European conference on speech communication and technology (Eurospeech-2005)*, 1585–1588. http://slp.itc.nagoya-u.ac.jp/web/papers/2005/eurospeech2005_tohyama_final.pdf . Accessed 13 Jan 2014.

Vardè, Sonia. 2014. *La smartpen per la didattica dell'interpretazione consecutiva*. Unpublished MA Thesis, Advanced School of Modern Languages for Translators and Interpreters (SSLMIT), University of Trieste.

Voinova, Tanya, and Noam Ordan. 2016. Narratives of community interpreters: What can we learn from using corpus-based methodology? In *Addressing methodological challenges in*

Interpreting Studies research, ed. C. Bendazzoli, and C. Monacelli, 107–139. Newcastle upon Tyne: Cambridge Scholars Publishing.

Vuorikoski, Anna-Riitta. 2004. *A voice of its citizens or a modern tower of Babel? The quality of interpreting as a function of political rhetoric in the European Parliament*. Tampere: Tampere University Press.

Wang, Binhua. 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, Binhua. 2012b. Interpreting strategies in real-life interpreting: Corpus-based description of seven professional interpreters' performance. *Translation Journal* 16 (2). http://translationjournal.net/journal/60interpreting.htm . Accessed 28 Sept 2016.

Wang, Binhua. 2015. Corpus-based interpreting studies in China. Paper presented at the conference corpus-based interpreting studies: The State of the Art. First Forlì International Workshop. University of Bologna at Forlì, Italy 7–8 May 2015.

Zagrebelsky Prat, Maria Teresa (ed.). 2004. *Computer learner corpora. Theoretical issues and empirical case studies of Italian advanced EFL learners' interlanguage*. Alessandria: Edizioni dell'Orso.

Zanettin, Federico, Silvia Bernardini, and Dominic Stewart (eds.). 2003. *Corpora in translator education*. Manchester: St Jerome.

Zanettin, Federico, Gabriela Saldanha, and Sue-Anne Harding. 2015. Sketching landscapes in translation studies. A bibliographic study. *Perspectives: Studies in Translatology* 23 (2): 1–22.

New Frontiers in Translation Studies

Mariachiara Russo
Claudio Bendazzoli
Bart Defrancq *Editors*

Making Way in Corpus-based Interpreting Studies

Springer

New Frontiers in Translation Studies

Series editor

Defeng Li
Centre for Translation Studies, SOAS, University of London, London, United Kingdom,

Centre for Studies of Translation, Interpreting and Cognition, University of Macau, Macau SAR

More information about this series at http://www.springer.com/series/11894

Mariachiara Russo · Claudio Bendazzoli
Bart Defrancq
Editors

Making Way in Corpus-based Interpreting Studies

Springer

Editors
Mariachiara Russo
Department of Interpreting and Translation
University of Bologna
Forlì
Italy

Claudio Bendazzoli
Department of Economic and Social Studies, Mathematics and Statistics
University of Turin
Torino
Italy

Bart Defrancq
Department of Translation, Interpreting and Communication
Ghent University
Ghent
Belgium

ISSN 2197-8689 ISSN 2197-8697 (electronic)
New Frontiers in Translation Studies
ISBN 978-981-10-6198-1 ISBN 978-981-10-6199-8 (eBook)
https://doi.org/10.1007/978-981-10-6199-8

Library of Congress Control Number: 2017950038

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

In memory of Miriam Shlesinger
An invaluable source of inspiration to us all

Foreword

Two decades have passed since Miriam Shlesinger put forward the idea of creating a new research domain within the discipline of Translation Studies in her inspirational paper entitled "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies", published in a Special Issue of *Meta* devoted to the Corpus-based Approach (Vol. 43 n. 4, 1998). Today CIS is undoubtedly a burgeoning area of scholarly inquiry and I am honoured and humbled to have been invited to write a foreword to the present collected volume, which testifies to the breadth and depth of knowledge gained from the systematic study of interpreting through corpora. In line with the founder of CIS, interpreting is here intended as "the production of oral output based on other-language input which may be either written (to be read) or unwritten (impromptu)" (Shlesinger 1998: 486–487). Fully recognizing the lasting value of Shlesinger's foresight, I wish to offer a reading of the papers published in this volume through the lens of her first contribution to the field with a view to assessing the advances made to date towards developing "a full-fledged paradigm of corpus-based interpreting studies" (1998: 487). When reading each chapter in turn, one notices immediately that Descriptive CIS takes centre stage. Inspired by Shlesinger's initial proposals, the contributors focus on the specificity of interpreting *qua* interpreting as a form of mediated oral discourse vis-à-vis original oral discourse and non-native language use. Cross-linguistic differences are also investigated and given due consideration when accounting for the linguistic patterns that characterize the interpreters' target speech. Moreover, different kinds of interpreting are examined besides simultaneous interpreting. Corpus typology has expanded considerably with the creation of monolingual comparable corpora as well as bilingual and multilingual parallel corpora in a variety of language pairs. The theoretical frameworks the present scholars draw on are also diverse and encompass not only cognitive psychology, as "[o]ne of the many paradigms which hold promise for the study of interpreting" (Shlesinger 1998: 489), but also gender studies, contrastive linguistics and media studies. Furthermore, the object of investigation is as varied as figurative language, interpreting strategies and norms and gender-based features. Also, the link between descriptive and applied CIS is implicit in most papers since the insights yielded by

these well designed empirical studies may well stimulate teachers to formulate and apply some "bridging rules" (Toury 2012: 11) that will enable interpreter trainees to appreciate the interrelationship between theory and practice and comply to set norms in a conscious way. Many of the methodological challenges discussed by Shlesinger have been met successfully. Others still remain, as is clearly acknowledged by all the present authors. Nonetheless, I can affirm with reasonable confidence that, thanks to scholarly endeavours such as the ones reported in this commendable publication, the benefits deriving from applying a corpus-based methodology to the study of interpreting far outweigh the difficulties. And Shlesinger's vision, beyond any shadow of doubt, has to date developed more than she had ever thought possible.

Sara Laviosa
University of Bari, Italy

References

Shlesinger, Miriam. 1998. Corpus-based Interpreting Studies as an offshoot of corpus-based Translation Studies. *Meta* 43(4): 486–493.
Toury, Gideon. 2012. *Descriptive Translation Studies—and beyond*. Revised Edition. Amsterdam/Philadelphia: John Benjamins.

Preface

The idea of editing a volume entirely focused on corpus-based interpreting studies was first discussed following the *First Forlì International Workshop on Corpus-based Interpreting Studies: The State of the Art*, which was held at the Forlì Campus of the University of Bologna on May 7–8, 2015. This event gathered more than 100 scholars from different parts of the world with the aim of sharing their corpus-based research endeavors, ranging from studies that exploited fully machine-readable corpora to small collections of texts or transcripts for manual analysis. The workshop came after other occasions in which translation and interpreting scholars presented their corpus-based research projects, though to differing extents. In 2010, a full session on "Interpreting Corpora" was included in the program of the *Emerging Topics in Translation and Interpreting—Nuovi percorsi in traduzione e interpretazione* conference organized by the University of Trieste (Straniero Sergio and Falbo 2012). The 2003 conference held in Pretoria *Corpus-based Translation Studies: Research and Applications* was probably the first one with an exclusive focus on corpus-based research in translation, though no contributions from interpreting scholars were presented. The volume that originated from that conference (Kruger et al. 2011) does include a chapter on CIS anyway, based on a paper presented at the 2006 IATIS conference (Bendazzoli et al. 2011). Going backward in time, it is clear that corpus-based interpreting studies, as an "offshoot" of corpus-based translation studies have flourished considerably and can be expected to develop even further.

This volume serves a dual purpose. On one hand, it aims at promoting the understanding of the interpretation process and product based not on anecdotal observations or small-size case-studies, but on comparatively large datasets of professional interpretations mostly stored and queried according to standard corpus linguistics methodologies. The volume showcases descriptions of and studies on major interpreting corpora available to date: the EPIC Corpus and its off-springs EPTIC (including also translations) developed at the University of Bologna, EPICG from the University of Ghent (Belgium) and the TIC Corpus from the University of Poznán (Poland); the 2249i Corpus, the DIRSI Corpus and the IMITES Corpus, again from the University of Bologna (Italy); the CorIT Corpus from the University

of Trieste (Italy); the FOOTIE Corpus created at UNINT University in Rome (Italy); the NAIST Corpus from the Nara Institute of Science and Technology (Japan) and the CEIPPC Corpus, which was built at the Guangdong University of Foreign Studies (China). On the other hand, the volume is also intended as a renewed call (after Miriam Shlesinger's first call in 1998) to the research community to further develop the field of corpus-based interpreting studies by offering scholars more corpus-based data and methodologies to compile their own corpora according to their research designs.

This volume consists of 11 chapters. The first two are meant to provide the theoretical framework of corpus-based interpreting studies (CIS), focusing on their development and on procedural-methodological issues respectively, while the other nine chapters present the insightful results obtained by analyzing professional interpreters' performances from this promising corpus-based perspective.

What follows is an overview of the contributions to the volume.

The book opens with the chapter "Corpus-based Interpreting Studies: Past, Present and Future Developments of a (wired) Cottage Industry" by Claudio Bendazzoli, who takes stock of nearly two decades of corpus-based studies all the way up to Web 2.0 applications and looks at the extent to which they have differently flourished in three areas of interpreting, namely research, education, and professional practice.

This is followed by a collective chapter "Building Interpreting and Intermodal Corpora: A *How-to* for a Formidable Task" by Silvia Bernardini, Adriano Ferraresi, Mariachiara Russo, Camille Collard, and Bart Defrancq, who pooled their pioneering experiences to provide an accessible step-by-step guide for corpus developers, especially those who are working with European Parliament (EP) data, and an appraisal of available technologies to cater for different research questions. The ultimate goal is to harmonize procedures in order to expand EP interpreting and multimodal corpora through a collective effort.

An example of corpus-based study is offered by Bart Defrancq and Koen Plevoets' "Over-uh-load, Filled Pauses in Compounds as a Signal of Cognitive Load", which opens the series of the chapters investigating interpreter strategies during simultaneous interpreting (SI). Based on their EPICG Corpus, a French-Dutch-English-Spanish corpus, the authors contrast SI data in Dutch with non-mediated Dutch linguistic production to ascertain the increased cognitive load associated with the production of compound lexemes between languages with compound parts in reverse order *versus* the source language (French).

A similar research question was investigated by Binhua Wang and Bing Zou in "Exploring Language Specificity as a Variable in Chinese-English Interpreting. A Corpus-based Investigation". Based on the Chinese-English Interpreting for Premier Press Conferences Corpus (CEIPPC), the authors studied the cognitive load associated with interpreting in the consecutive mode between two languages with major differences in cultural conceptualizations and linguistic structures, Chinese and English. In particular, they focus on the processing of the following asymmetry:

attributive modifying structures which are typically front-loaded in Chinese and modifying structures which are typically back-loaded in English.

The language of the professional interpreter or *interpretese* is the subject of study of the two following chapters. Guy Aston in his contribution "Acquiring the Language of Interpreters: A Corpus-based Approach" discusses the value of memorized formulae to produce fluent speech and, based on his 2249i Corpus, stresses the potential of corpora to detect fixed expressions in professional simultaneous interpreters and store them for the benefit of trainee interpreters. Marta Kajzer-Wietrzny's chapter "Interpretese vs. Non-native Language Use: The Case of Optional *That*" describes the functions of the optional complementizer *that* and compares its use and the *zero* variant in interpreted, non-native and native English discourse at the European Parliament, collected in her TIC Corpus, in order to detect the prevailing linguistic patterns among simultaneous interpreters.

A novel field of research in corpus-based interpreting studies concerns gender. Mariachiara Russo's chapter "Speaking Patterns and Gender in the European Parliament Interpreting Corpus" investigates a number of parameters, i.e. speaker's mode of delivery, input speed, language combination, and topic, in relation to target speech length revealing some statistically significant differences among female and male interpreters.

The subsequent three chapters analyze simultaneous interpreters' strategies when dealing with very challenging speech acts: the use of figurative language by the speaker and fast adversarial exchanges during a political debate and a press conference.

In her contribution "Studying Figurative Language in Simultaneous Interpreting: The IMITES (*Interpretación de la Metáfora entre ITaliano y ESpañol*) Corpus", Nicoletta Spinolo classifies the linguistic behavior of interpreters faced with 1135 figurative expressions and identifies the nature of those most difficult to translate.

Eugenia Dal Fovo's study "European Union Politics Interpreted on Screen: A Corpus-based Investigation on the Interpretation of the Third 2014 EU Presidential Debate" is based on her EUDEB14 Corpus, a subcorpus of CorIT, the world largest TV interpreting corpus developed by our late colleague and friend Francesco Straniero Sergio at the University of Trieste. The author contrasts interpreting norms and ethics between SCIC interpreters and free-lance interpreters for the same event, the debate among candidates to the Presidency of the EU Commission, to ascertain to what extent the composition of the interpreting team and the equipment influence the representation of the interaction. Her results reveal that the freelance interpreters, who were TV interpreting experts displayed a more telegenic style, in keeping with the spectacularization principle and the confrontational dynamics of televised political debates.

Interactional dynamics were also analyzed by Annalisa Sandrelli in a totally different setting. In her chapter, "Interpreter-mediated Football Press Conferences: A Study on the Questioning and Answering Strategies", she observes how the target language versions closely mirrored the source language Q&A functions, despite the

fast changes in turns at talk, overlapping speech and the psychological pressure on the interviewees.

The eleventh and last chapter adds another perspective to corpus-based interpreting studies, that of technologists Graham Neubig, Hiroaki Shimizu, Sakriani Sakti, Satoshi Nakamura, and Tomoki Toda who are interested in understanding the difficulties faced by human interpreters and the possibilities of creating systems that help interpreters overcome these difficulties through the creation of assistance tools or speech translation (ST) technology. Based on the NAIST Japanese-English Corpus, they compare the interpreted output of interpreters of varying degrees of experience with the translated output of the same source speeches. Their chapter "The NAIST Simultaneous Translation Corpus" describes the collection of source language materials, interpretation processes, recording, and transcript of resulting data.

We hope that the richness of approaches and results offered by the present volume may inspire other scholars to join efforts and resources with a view to expand interpreting corpora and validate interpreting hypotheses on larger datasets. Likewise, we hope that also trainee interpreters may benefit from being exposed to such a wide range and abundance of professional interpreting styles and successful strategies.

Forlì, Italy — Mariachiara Russo
Torino, Italy — Claudio Bendazzoli
Ghent, Belgium — Bart Defrancq

References

Kruger, Alet, Kim Wallmach, and Jeremy Munday. eds. 2011. *Corpus-based Translation Studies: Research and applications*, London/New York, Continuum.

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

Straniero Sergio, Francesco, and Caterina Falbo. 2012. *Breaking ground in corpus-based Interpreting Studies*. Bern: Peter Lang.

Contents