The International Journal for Translation & Interpreting Research
trans-int.org

Book Review

Francesco Straniero, S. & C. Falbo (Eds.) (2012) *Breaking Ground in Corpus-based Interpreting Studies*. Linguistic Insights Series 147. Bern: Peter Lang AG

Reviewed by Fang Tang
Guangdong University of Foreign Studies, China
candy.fangtang@hotmail.com

DOI: ti.106201.2014.r01

Mona Baker's seminal article "Corpus Linguistics and Translation Studies — Implications and Applications" (1993) has been widely recognized as the manifesto kicking off Corpus-based Translation Studies (CTS). Since then, the methodology supporting the development of Corpus Linguistics has been applied frequently to Translation Studies (TS). It has even set off, in Laviosa's words, a "corpus linguistics turn in TS" (Laviosa 2004: 29). This paradigm shift echoes with Toury's target-text-oriented perspective in facilitating the search for "the nature of translated texts as mediated communicative events" (Baker 1993: 263).

Interpreting studies, as a sub-branch of Translation Studies, gain momentum in cooperating with Corpus Linguistics mainly from Shlesinger, who in her 1998 article first gives a full account of problems and benefits of Corpus-based Interpreting Studies (CIS) and later in her 2008 article reports findings of stylistic and pragmatic features of interpreted texts as compared with translated texts. Yet, as Thompson summarizes "The recording and transcription of unscripted speech events is highly labor intensive in comparison to the work involved in collecting quantities of written text for analysis" (2005: 254); during these years, it is always the accessibility of interpreted texts and their inherently complex transcription process that impede the development of CIS.

As a breath of fresh air in Interpreting Studies, this collection is the first book devoted especially to CIS. In general, it provides readers with a comprehensive view of five heterogeneous interpretation corpora (EPIC, DIRSI, CorIT, FOOTIE, and one corpus constituted by court hearings) and various research efforts originating from them. These five corpora are heterogeneous because each corpus is situated in different locations: EPIC represents SI in political sessions; DIRSI concerns about SI in medical conferences; CorIT relates to SI and CI for TV program; FOOTIE is relevant to in sports settings; the last corpus is pertinent to dialogue interpreting inside courtrooms.

This book starts with an exhaustive introduction sketched by its two editors Francesco Sergio and Caterina Falbo. They help at the very outset by presenting a detailed literature review of CTS, which sets a solid theoretical stage for the articles that follow. The aim of CTS has been generalized as the quest for “the nature of translated text as a mediated communicative event” (Baker 1993: 243), translation universals, laws and norms. Two features of CTS (also applicable to CIS) have been revealed: 1) linguistic features of translated language can be put into two categories: one is attributed to cognitive factors and can be identified through process-oriented research while the other is ascribed to the social, historic and cultural factors and can be investigated through product-oriented research; 2) findings of CTS are descriptive, only reflecting how translators translate; reasons lying behind choices still need to be explored by researchers.

A series of criteria defining the representativeness of interpretation corpus are outlined, which claim that choices among parameters like interpreter’s expertise level, sex, age, situational context (real/experimental; TV/telephone/court/ medical, etc.), interpreting mode (simultaneous/consecutive/liaison, etc.), language and directionality should be subject to the objective of the study.

In the second article, Mariachiara Russo and Claudio Bendazzoli introduce the first corpus EPIC (European Parliament Interpreting Corpus), where several European Parliament plenary sittings held in 2004 (including February, March, April and July) are recorded with authorization. It comprises nine sub-corpora, including source speeches in Italian, English and Spanish and interpreted speeches from English to Italian and Spanish, from Italian into English and Spanish, as well as from Spanish into Italian and English. This complex structure facilitates both comparable and parallel analysis.

To make the audio or video output machine-readable, a “labor-intensive and arduous” transcription process is inherently needed (Shlesinger, 1998: 2). The transcription of EPIC has been facilitated by EU officials who have done verbatim reports and translation of the EP plenary debates and by speech recognition programs (Dragon Naturally Speaking and Via Voice) which speed up the transcription of targeted speeches. Since “certain elements of spoken communication are both so subtle and so subjective” (Cook 1995: 51-52; O’Connell et al. 1993; cited from Shlesinger 1998: 2), not all characteristics of the original speech can be reproduced in written form. Thus, transcription turns to be a selective process, where features like pauses, repetitions, prosody and body language may be deleted based on the nature of the material and the aim of the research. For EPIC, only basic transcripts are provided: for the linguistic information, there is no punctuation and units of meaning are segmented based on the speaker’s intonation and syntactic information available. The end of each segment is indicated by the double bar sign (//), which would facilitate the alignment process between source and target speeches; for the paralinguistic information, only truncated words, mispronounced words, pauses are presented; for extra-linguistic information, a specially-designed header showing information like context, speaker, duration, topic, etc. are provided so as to ensure automatic queries. The transcripts have been POS-tagged (Italian and English speeches are done by Treetagger while Spanish ones are by Freeling) and lemmatized. It should be noted that the POS-tagging process may sometimes be misled by repetitions, ungrammatical structures, interjections, the absence of punctuation and the high number of neologisms, technical terms and EU-jargon words, etc. To facilitate query, the tagged output is converted to .xml format and indexed by using the IMS corpus Work Bench – CWB. The web interface and query tools of EPIC have also been graphically presented.

So far, several studies have been conducted based on EPIC (Russo et al. 2006; Sandrelli et al. 2010). Russo et al. (2006) touched on questions like whether interpreted texts have a lower lexical density and less lexical variety than original texts and whether lexical patterns change according to language pair and language direction. Results suggest that the interpreted texts show: 1) a trend of text compression; 2) a higher lexical density than originals in the same language, which is the opposite of what Laviosa (1998) has found in relation to translated texts; 3) a lower lexical variety than originals in the same language, which is in line with Laviosa's (ibid.) finding on translational English.

Disfluencies have also been looked into, based on the marked truncated and mispronounced words (Sandrelli et al. 2007; Bendazzoli et al. 2011). Results suggest that truncated words are a much more frequent type of disfluency than mispronounced words in all three languages. They are frequently repaired not only by speakers but interpreters (evidence of self-monitoring mechanism). Overall, both types of disfluency are more frequent in interpreted speeches than in speeches originally delivered, confirming cognitive demands of SI may lower the efficiency of self-monitoring.

Assumptions on whether disfluencies vary with topics, mode of delivery and speed have been examined and results reflect that: 1) when the topic is "procedures and formalities", few disfluencies can be identified, which might be attributed to interpreters' familiarity with the formulaic language and parliamentary routines as well as the availability of agendas and lists of speaker; 2) when the original is an impromptu speech, few disfluencies can be identified, which might be ascribed to its relatively lower information density than read speech; 3) no conclusive result can be identified about the interaction between frequency of disfluencies and the varying speed of the originals, probably because EP interpreters have been accustomed to working at high input rates.

To find out whether cognate languages may pose interferences in SI, morpho-syntactic asymmetries and lexical/syntactical ambiguities are automatically extracted from EPIC (speeches from Spanish into Italian) through multi-item searches. Results indicate that the source language asymmetries and ambiguities in Spanish source speeches exert only small impact on relevant EP interpreters and these SL items does not seem to affect their ST comprehension.

The EPIC multimedia archive has even fostered several dissertations, whose topics range from grammar, through interpreting strategies to pragmatics issues.

EPIC is under continuous development: 1) its size has been expanded; to balance the distribution between languages, more Spanish and Italian originals and their interpretations will be added; 2) text-video alignment for SL speeches and text-sound alignment for interpreted speeches are made through SpeechIndexer and Transana 2.41.

The second corpus DIRSI-C (Directionality in Simultaneous Interpreting Corpus), as illustrated by Bendazzoli Claudio, is a bilingual (Italian and English) speech corpus comprising recordings and transcripts from three international medical conferences. Its theoretical and methodological framework is based on that of EPIC. DIRSI-C is both a comparable and parallel corpus with four subcorpora: one Italian original speech and its simultaneous interpretation into English as well as one English original speech and its simultaneous interpretation into Italian. It altogether has nearly 136,000 words, with a balanced distribution among the four subcorpora.

Source speeches are recorded with a laptop connected to the floor sound system of the conference halls and edited by sound editor software CoolEdit Pro while interpretations are recorded by either a micro digital recorder or another laptop in the booth. All the recordings are saved in .wav format. The concerned interpreters are four Italian native speakers and one English native speaker. Its transcription method is similar to that of EPIC. The STs have been further classified according to their duration (short, medium, long), length (short, medium, long) and speed (low, medium, high). Based on theories from Linguistic Anthropology, the Ethnography of Speaking, Sociolinguistics, Conversation Analysis and Discourse Analysis, Bendazzoli has improved approaches to study human communication and modified analytical tools through a refined taxonomy of the header of each speech in the corpus, i.e. speech event and participants.

The third corpus is FOOTIE (Football in Europe) created by Annalisa Sandrelli. The data are collected during the press conference of the 2008 European football championship, involving SI from Italy to Holland, to Romania, to France and to Spain. This corpus is marked as being both parallel and comparable, multimedia (with audio recordings of pre-match conferences and video recordings of post-match recordings and written transcripts of both), multilingual (involving Italian, English, French and Spanish), single-genre (only dialogue), synchronic (all happened in June 2008), closed (only with 16 EURO 2008 press conferences) and untagged.

FOOTIE demonstrates typical cases of institutional interaction, where communicative roles are pre-determined and mutually exclusive. For instance, journalists are responsible for raising questions while protagonists are obliged to answer them; journalists are the only audience on the spot because the press conferences are not broadcast live by European TV channels for the general public but just to give journalists access to the latest information on games through interviewing protagonists. So interaction in this special occasion comes to the forefront of Sandrelli's research. Unlike other researchers who focus only on primary participants, Sandrelli gives snapshots of every participant involved, i.e. overhearers like volunteers, moderator like chairs or discussants. In her conclusion, she summarizes interesting interaction features: usually the floor allocation speech events in the corpus have not been translated but turned into an opportunity for interpreters to relax; Since journalists' questions always start with a new topic and are relatively short, interpreters' EVS should be kept very short to catch up with the beginning of following questions; interpreters may initiate a turn to remind participants of some technical problems; interpreters produce a faster output when interpreting from A to B language while slow down their delivery when interpreting from B to A language.

CorIT (an Italian Television Interpreting Corpus), the fourth corpus in this volume, is built by Caterina Falbo. CorIT is the only diachronic corpus in this collection (as it has recorded interpreting's appearance on Italian television for almost fifty years, starting from the first SI during the Moon landing coverage to the latest US presidential debates and British prime ministerial debate). It also features as being open (as it has been expanded continuously and now includes more than 2,700 interpretations), multimode (including both CI and SI), unidirectional (with interpretations from various foreign languages into Italian), multimedia (including video recordings and transcripts) and partially parallel (due to the partial unavailability of original texts). To optimize the search function, each interpretation has been provided with an individual code to represent information like interpreter's name, interpreting mode, interaction type, name of the participants in the communication event, date, Italian program/broadcasting channel, foreign broadcasting channel, macro-genre, television genre, type of text, etc. WinPitch is the only software adopted for transcribing due to its ability to slow down the original recording without distorting it, to highlight overlaps and to conform to technical and ergonomic characteristics. Special rules for transcription have also been expounded, i.e. punctuation should be omitted;

speakers should be identified in the transcript; doubts encountered by the transcriber should be marked, etc.

The succeeding two essays report two research projects based on CorIT. The first is from Eugenia Fova who discusses topical coherence in the SI of the question and answer part (Q&A) of American presidential debates broadcast on Italian television between 1988 and 2004 in both her MA and PhD studies. In her MA thesis, Fovo puts the types of questions in the source and target texts into four categories: 1) yes/no questions; 2) wh- questions; 3) leading questions and 4) declarative questions. This typology system serves as a launching pad for the following SL-TT contrastive analysis, where each Q&A occurrence will be identified and information pertaining to its type of question, degree of coherence achieved in the TT, shifts made by the interpreter, and whether structure is maintained are analyzed. Results reveal that: 1) the most frequently recurring question is the Wh-type (57%); 2) the most frequently omitted question is the Y/N type; 3) when the question's structure is maintained, shifts (such as omissions or substitutions) do not necessarily result in unsuccessful rendition or lack of coherence; whereas when interrogative-clause types is changed, even though no shift has been made and the segment seems cohesive, a lack of topical coherence is very likely to occur.

Fova's on-going doctoral research project is a further study of coherence. To remedy shortcomings in her MA thesis, two critical theoretical issues have been refined: the identification of topical coherence and the definition of the Q&A group. For the first issue, a communication framework has been defined after a review of studies on coherence in interaction from Text Linguistics as well as Interaction and Conversation Analysis; for the latter issue, to devise a complete analysis grid, interpreted questions will on the one hand be investigated as an autonomous section in terms of type, structure and internal topical coherence and on the other hand be observed through an OT (original text) / IT (interpreted text) contrastive analysis.

In the seventh article, Francesco Sergio provides an interesting investigation of individual interpreter's interpreting style. This is a field rarely discussed (till now only two relevant studies have been found by Sergio). After an introduction of previous studies, Sergio lists a few elements that can be served as examples of interpreter's style, such as the use of the adjective "straordinari", turning the use of lexical couplets into a translational habit, the aversion of using simple and direct equivalents and favoring of equivalents which change the illocutionary force of the original utterances.

Sergio draws his data from one of the CorIT subcorpora, namely Grand Prix Formula One Press Conferences, which cover 340 press conferences interpreted by 26 interpreters from 1997 to 2010 (the SI output being 30 hours or so). He focuses on four major interpreters and investigates not only their type token ratio but also the first ten favorite DMs (discourse markers) they adopted.

Sergio also conducts a case study on Olga Fernando, Italy's most popular media interpreter with over 20 years' experience of TV interpreting. Two typical features have been unveiled: 1) different from most TV interpreters who follow closely behind the speaker and try to reduce concurrent speaking and listening, Ms. Fernando features with long décalage in SI. She spends more time in listening and tends not to reproduce before acquiring meaningful segments; 2) in CI, she often starts after the speaker finishes his/her turn, so as to ensure fluent reproduction.

The challenge that researchers may encounter while collecting data in courtrooms is given its fullest and most focused treatment in the final article, where Marta Biagini elucidates three major challenges she encounters: the first one is decision on the language combination. Her study is with Italian-French combination due to her familiarity with them; the second challenge concerns applications for permission and selection of proceedings: Researchers in Italy should ask permission from the *Presidenti* of relevant Italian courtrooms. Yet only certain court can be recorded for confidentiality consideration; for the third issue of recordings, two ways have been offered: one is obtaining audio-recordings from the court's archives and the other is using the researcher's own recordings. Currently, Biagini's corpus consists of about nine hours of recordings of French-Italian dialogue interpreting contributed by six interpreters. Data has been transcribed using the software WinPitch and following conventions adopted by French researchers in spoken language and oral verbal interactions.

To conclude, as one of the few books with an exclusive focus on the slowly-developed subject of CIS, it can definitely be regarded as a welcome addition to the repertoire of interpreting studies literature. Essays in the volume provide readers with an inspiring account of not only operational steps to design and build interpreting corpus but also possible approaches to implement relevant investigations in some way or another. Its highly informative nature has surely set this volume as a landmark to stimulate further corpus-based exploration on features of what Shlesinger called "interpretese" (2008: 237).

References

Baker, M. (1993) Corpus Linguistics and Translation Studies: implications and applications. In Baker, M., Francis, G., & Tognini-Bonelli, E. (eds.) *Text and Technology: In Honor of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 233-250.

Bendazzoli, C., Sandrelli, A., & Russo, M. (2011) Disfluencies in simultaneous interpreting: a corpus-based analysis. In Kruger, A., Walmach, K., & Munday, J. (eds.) *Corpus-based Translation Studies: Research and Applications*. London/New York: Continuum, 282-306.

Laviosa, S. (1998) The Corpus-based approach: a new paradigm in Translation Studies. *Meta,* 43 (4): 474-479.

Laviosa, S. (2004) Corpus-based Translation Studies: Where does it come from? Where is it going? *TradTerm,* 10: 29-57.

Russo, M., Bendazzoli, C., & Sandrelli, A. (2006) Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: extended analysis of EPIC (European Parliament Interpreting Corpus). *Forum,* 4(1): 221-254.

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

Shlesinger, M. (2008) Towards a definition of Interpretese: An intermodal, corpus-based study. In Hansen, G., Chesterman, A., & Gerzymisch-Arbogast, H. (eds.) *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile*. Amsterdam/Philadelphia: John Benjamins, 237-254.

Sandrelli, A., Bendazzoli, C., & Russo, M,. (2010) European Parliament Interpreting Corpus (EPIC): methodological issues and preliminary results on lexical patterns in simultaneous interpreting. *IJT – International Journal of Translation,* 22: 165-203.

Sandrelli, A., Russo, M. & Bendazzoli, C. (2007) *The impact of topic, mode and speed of delivery on interpreter's performance: a corpus-based quality evaluation*. Poster presented at the International Conference Critical Link 5. Quality in Interpreting: A Shared Responsibility, 11-15 April 2007 Sydney, Australia.

Thompson, P. (2005) Spoken language corpora. In Wynne, M. (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 59-70.