Chapter 2
Building Interpreting and Intermodal Corpora: A *How-to* for a Formidable Task

Silvia Bernardini, Adriano Ferraresi, Mariachiara Russo, Camille Collard and Bart Defrancq

Abstract This contribution has a double aim. On the one hand, it highlights the various challenges and problems compilers of (simultaneous) interpreting and intermodal corpora are likely to face, and the solutions that were found and applied in three corpora of European Parliament plenary debates, i.e. EPIC, EPICG and EPTIC. On the other, it provides an accessible step-by-step guide for corpus developers who are working with European Parliament data, with the ultimate aim of bringing as far as possible into line the procedures used to transcribe the audio tracks, record metadata, annotate texts with part-of-speech and lemma information, perform text-to-text and text-to-audio/video alignment, and index the corpus for searching with appropriate corpus query tools. By adopting shared corpus building methods, it might be possible to leverage the substantial efforts already deployed by different research groups, and encourage others to take charge of new language pairs. This, we shall argue, might lead to a massively multilingual interpreting and intermodal corpus, through a distributed community effort.

Keywords European Parliament data · Interpreting corpora · Intermodal corpora Transcription of oral data · Corpus annotation

S. Bernardini (✉)
Department of Interpreting and Translation, University of Bologna,
Corso della Repubblica 136, 47121 Forlì, Italy
e-mail: silvia.bernardini@unibo.it

A. Ferraresi · M. Russo
Department of Interpreting and Translation,
University of Bologna, Forlì, Italy

C. Collard · B. Defrancq
EQTIS, Ghent University, Ghent, Belgium

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_2

2.1 Introduction

Readers of this volume will no doubt concur that interpreting corpora are valuable resources, that lend themselves to a host of investigations and practical applications, such as those surveyed in the first chapter in this volume. Yet very few such corpora exist, fewer still are publicly available, and all of them are, by current standards, microscopic. This dearth is hardly surprising, since interpreting combines two features that have traditionally hindered the development of corpus resources: orality and interlinguistic mediation. As concerns the former, speech, particularly if impromptu, has been suggested to be among "the most difficult and expensive [language varieties] to acquire, difficult to classify and manage" (Sinclair 1996, unpaginated). As for the latter, complexity is inherent in the nature of corpus-based translation research, which "is always based on a comparison between corpora of different types so that, in translation studies, a corpus is actually always a combination of at least two subcorpora" (Zanettin 2013, p. 26).

The combination of features of orality and features of mediation makes interpreting corpora complex research constructs, for which attaining "maximum usability, reliability and "longevity"" (Ruhi et al. 2014, p. 1, quotes as in the original) is both critical and extremely challenging. At the same time, however, there is no denying that "spoken corpora users and developers have […] their specific research goals and may "cherish" their own ways of doing things" (ibid., p. 2, quotes as in the original). Interpreting corpora have recently started to be developed also as components of intermodal corpora, i.e. corpora which bring together different mediation modes (written translation *vs*. spoken interpreting; Shlesinger 2009). The expectations and research priorities of developers in this case may be different from those of "pure" interpreting scholars, and may result in corpus building decisions that effectively thwart chances of uptake and further development by different sets of potential stakeholders.

One of the most promising sources of interpreting and intermodal corpora is no doubt the European Parliament (EP). The availability of interpretations and translations from and into a large number of languages, the ease of access to the videos (downloadable from the Internet), and the high professional standards of the interpreters and translators involved, make the EP a dream setting from which to draw a very large, multilingual, intermodal corpus. And indeed, various initiatives are currently underway to create EP corpora that include a simultaneous interpreting component (notably in Bologna, followed by Ghent, Belgrade, Louvain, Lisbon, Poznan and Saarbrücken). Different languages are being covered (so far: English, French, German, Dutch, Italian and Spanish as sources and targets, with Finnish and Slovenian under construction), and published verbatim reports of EP speeches with their translations are being collected for some language pairs, so as to move towards fully intermodal corpora of interpretations and translations. In this contribution, we refer to this set of EP corpora as "the EPIC suite of corpora", to acknowledge the pioneering role played by the creators of EPIC, the European Parliament Interpreting Corpus (Russo et al. 2006), with regards to EP corpora.

The remainder of the chapter discusses the challenges and problems that compilers of (simultaneous) interpreting and intermodal corpora are likely to face, and present the solutions adopted for the EPIC suite of corpora. Specifically, Sect. 2.2 is devoted to corpus design principles, with special emphasis on the transcription of oral data and the collection of metadata. Section 2.3 examines annotation issues related to linguistic tagging and text-to-text and text-to-audio/video alignment, and Sect. 2.4 discusses how to integrate texts and different layers of annotation so that they can be profitably consulted through a corpus query tool. Taken together, the indications provided throughout the chapter should also constitute an accessible step-by-step guide for corpus developers working with EP data. As suggested in Sect. 2.5, we thus hope to encourage research groups to join efforts, leading to the construction of a massively multilingual interpreting and intermodal corpus.

2.2 Corpus Design and Compilation

2.2.1 The Basics

How a corpus is designed and compiled ultimately depends on, and at the same time constrains, what it will be used for. The minimal option is to transcribe interpretations and to assemble them in a collection of searchable files. Such a corpus can then be exploited for research into the specific features of interpreted language, in comparison, for instance, with non-interpreted spoken language. This monolingual comparable approach is discussed in Shlesinger (1998) and used, e.g. in Russo et al. (2006) and Kajzer-Wietrzny (2012).

When approaching most other topics related to interpreting studies, however, the requirement for the corpus to also include transcriptions of source speeches is almost unescapable. This applies among others to the study of interpreter strategies (anticipation, chunking, etc.) and the quality/accuracy of interpreting, which requires that one has access to the properties of the source text in the first place. EPIC and EPICG (EPIC-Ghent) are good examples of corpora of this type.

Finally, if one is interested in intermodal comparisons, the corpus will have to include interpretations and translations, preferably from closely comparable or quasi-parallel source texts. An example of such corpus type is EPTIC, the European Parliament Translation and Interpreting corpus.[1]

Transversally to the corpus types just surveyed, with the sociological turn in interpreting studies (Angelelli 2012), contextual metadata have become increasingly important, as they provide the necessary background information to approach

[1]Throughout this chapter we describe the most recent, trilingual version of EPTIC, containing EP speeches in English, French and Italian from 2011 (Bernardini et al. 2016a). The first version of the interpreting subcorpus of EPTIC (containing speeches from 2004; see Bernardini et al. 2016b) is more similar to EPIC, from which it was derived.

interpreting as a socially situated activity. These crucially depend on researchers' decisions and can include anything from speaker and interpreter gender to information on the interactional process.

In this Section, we discuss the two basic ingredients for compiling interpreting and intermodal corpora: transcriptions and metadata.

2.2.2 Transcribing Interpreting Data

There are various ways of transcribing oral data, depending on the priorities researchers have and the solutions they must find to a series of problems. In this Section we first discuss general issues, and then look at specific problems and solutions adopted for the transcription of the EPIC suite of corpora.

2.2.2.1 General Issues

The dilemma transcribers face is very similar to the one translators and interpreters face: reconciling accuracy with regard to the source and adequacy with regard to corpus users' needs. As is the case with translation and interpreting, accuracy when transcribing can only be partial: the complexity of the acoustic signal is such that no written representation can do it justice. For many types of research, it is undesirable to include too many properties of the acoustic signal: they divert scarce resources to aspects that will only rarely be investigated, while making the data cluttered and thus less usable for the research most scholars are likely to be interested in. This is the reason why interpreting corpora rarely include phonetic transcriptions.

However, most corpus compilers do occasionally include relevant phonetic and prosodic properties. Disfluencies such as mispronounced words, truncated words, self-corrections and (filled/unfilled) pauses are generally transcribed along with the more standard segments of speech. In the field of simultaneous interpreting, it is important to include such features, as they are generally considered signs of cognitive load (Plevoets and Defrancq 2016). Similarly, *lapsus linguae* tend to be included in one way or another, along with salient prosodic features such as rising and falling intonation, high-pitched voice or salient word stress.

Another area in which dilemma looms is interaction. This is the case when compiling community interpreting corpora, whose data are typically drawn from instances of interaction from healthcare, legal or any other public service settings. Interactional features add several layers of complexity to a corpus: apart from the need to select data from contexts with limited numbers of participants in order to be able to keep track of their roles, there is also the need to signal interactional features, such as turn-taking, overlapping speech, turn-yielding cues, etc. in the transcription, without compromising its readability and searchability.

Most available transcriptions of interpreter-mediated interaction use Jefferson's conventions (Jefferson 2004) or simplified versions of them to present interactional

features. Such features may seem less relevant in corpora of simultaneous interpreting as the interpreter *stricto sensu* does not interact with the speaker. However, even in monologic discourse such as is typical of the EP, it is not uncommon to come across interaction (for instance by the moderator) that transcribers have to account for. Furthermore, the extent to which simultaneous interpreters and speakers overlap and the time span between semantically equivalent segments in source and target texts (EVS or *décalage*) are relevant features that raise transcription issues not dissimilar from those typical of traditional interaction.

Finally, corpora of oral data also regularly include references to observable phenomena other than the recorded voices of the participants, such as laughter and background noises, or gaze orientation and gestures (in the case of dialogue interpreting corpora). Gaze and gestures are especially relevant to comprehensively study turn management in interpreter-mediated conversations (Davitti and Pasquandrea 2017), while background noises may contain valuable information about processes that influence or are influenced by the interaction, such as, for instance, typing of written records of the interaction (Komter 2006).

In the next Section, we zoom in on the transcription methods and conventions used in the development of the EPIC suite.

2.2.2.2 The EPIC Suite: Transcribing Simultaneous Interpretations

Since 2008, the plenary sittings and some of the committee sittings of the EP can be watched online through the website of the European Parliament. The more recent corpora in the EPIC suite, such as EPICG and EPTIC, have been transcribed on the basis of downloaded audio/video files of speeches and interpretations.[2] The original EPIC was based on plenary sessions recorded on videotapes from the European Union's televised information channel Europe by Satellite.

The data for the original EPIC were transcribed using a shadowing technique[3] and a speech recognition software. The automatic transcriptions carried out by the software were manually cross-checked at a later stage to produce a final version in *txt* format. EPICG uses the corpus software suite EXMARaLDA, also used for the CoSi[4] and DiK[5] corpora compiled at the University of Hamburg. The video files were downloaded from the EP website, processed so as to obtain paired source-target audio recordings (using Handbrake to extract the monolingual tracks

[2]The speeches and the associated verbatim reports can be accessed via the European Parliament web page (http://www.europarl.europa.eu/plenary/en/debates-video.html), which allows searching by parliamentary term, date(s) of the sittings, name of speaker and keywords.

[3]"Shadowing […] involves the immediate vocalization of auditorily presented stimuli, i.e. word-for-word repetition, *in the same language*, parrot-style of a message presented through headphones" (Lambert 1992, p. 17). This technique is usually employed at the beginning of interpreter training to develop dual-task skills i.e. listening and speaking at the same time.

[4]https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:cosi.

[5]http://www1.uni-hamburg.de/exmaralda/files/k2-dik/public/index.html.

from the multilingual recordings, and VLC and Reaper to reassemble them), which were then imported into EXMARaLDA to be transcribed. For the creation of EPTIC, the published verbatim reports of the original speeches and their translations were used as a basis for transcription, saving the transcribers some keyboarding time. However, substantial editing was needed to restore the markers of orality, particularly in the interpretations. Since the revised translations are provided instead of actual interpretation transcripts, reverting to the original interpretation can at times prove as taxing as starting from scratch.

In general terms, data were transcribed orthographically, using the spelling and capitalization conventions prescribed by the EU Interinstitutional Style Guide.[6] Prosodic information (such as heavy word stress or high-pitched voice) is not included in any of the corpora, except for the question mark signaling rising intonation in EPICG. Salient observable phenomena, such as laughter and background noises are also included. The many other decisions that had to be made when transcribing EP interpreting data are too numerous to be discussed extensively here. Therefore, we simply provide selected examples of problems we were faced with, and the solutions we adopted.

Decisions concerning punctuation differed substantially across the different teams. EPICG uses no punctuation, apart from the question mark between square brackets [?], which is occasionally used to mark rising intonation. Transcripts in EPIC and EPTIC are segmented in sentence-like units of meaning by a double slash sign [//], based on prosodic and syntactic information. This information is essential to perform source-target and intermodal alignment, as well as Part-Of-Speech (POS) tagging (see Sect. 2.3). EPTIC further includes sub-sentential punctuation marks inserted by the transcribers taking into account prosodic and syntactic cues.

Mispronunciations, including words interrupted by an empty or filled pause, are transcribed as such, but the words are also included in their normalized form. To avoid duplication (e.g. in word counts), and to optimize tagging accuracy, the normalized version is included in the running text and the word as it was pronounced is represented as a sort of attribute of the first (within slashes). EPICG has opted to include the "as-is" version in the running text, while EPIC and EPTIC have included the normalized version in the running text. Truncated words are recorded as such, including a special character signalling the truncation ('-' in EPIC/EPTIC, '/' in EPICG). Inaudible segments are marked with a special character or with a comment ('#' in EPTIC and EPIC, '[inaudible]' in EPICG). Both empty and filled pauses are included with the transcriptions. Empty pauses are signalled with brackets (in EPICG) or suspension points (in EPIC/EPTIC), while filled pauses are transcribed orthographically (*euh, ehm,...*). Since EPICG is transcribed in EXMARaLDA, it also provides information on pause length.

Finally, as is the case in almost all transcription systems of oral language, numerals are spelled as words instead of figures in EPIC and EPICG. The date *2006* appears as *two thousand and six* in EPIC and *two-thousand-and-six* in EPICG.

[6] http://publications.europa.eu/code/en/en-000100.htm.

Table 2.1 Overview of codes used for interactional and non-verbal acoustic features

Feature	EPIC	EPICG	EPTIC
Silent pause	…	((0,3))	…
Filled pause	ehm	euh, euhm	ehm
Mid-word pause	proposal /pro_posal/	spea/euh ker [speaker]	proposal /pro_posal/
Rising intonation	NA	[?]	?
Non-verbalized noises	NA	[laughter]	[applause]
Non-standard pronunciation	NA	report [repo:rt]	proposal /prooposal/
Inaudible segment	#	[inaudible]	#
Mispronunciation	Parlamento / Parlomento/	intergoration [interrogation]	Parlamento / Parlomento/
Truncated words	propo-	propo/	propo-
Ambiguity	NA	{ce qui|ce qu'il}	NA
Overlapping talk	NA	| to do < L2 > what|	NA
Sentence-like segments	//	NA	.

Transcribing numerals as figures would inevitably lead to misrepresentations in case of repairs, hesitations, etc. (for instance transcribing *two ehm thousand ehm four* as *2 ehm 1000 ehm 4*). EPTIC instead follows the language-specific conventions provided in the EU Interinstitutional Style Guide, in most cases spelling numbers as figures. This is necessary for comparison with translated texts, which follow the same conventions.

By way of conclusion, Table 2.1 provides the different codes used to transcribe different types of interactional and non-verbal acoustic features in the EPIC suite of corpora.

2.2.3 Recording Metadata

As stated by Wörner (2012, p. 383), metadata are "*Data about Data or Information about Data*." In other words, they are not simply *data*, but

> […] structured data that describes *data resources* (in our case *language resources*), providing information about certain aspects of these resources (like contents or context) that add to the overall quality of the resource and makes it more accessible "to allow a better and more precise retrieval" (MetaGuide 2003). (Wörner 2012, p. 383)

EPIC/EPTIC metadata are included in a header for each transcript and fulfil a double purpose. On the one hand, they provide relevant contextual information on

the oral data considered as situated speeches—i.e. speech events delivered by specific speakers in a given context, which allows for investigations in keeping with the standards of ethnography of communication. On the other hand, they make it possible to restrict queries based on structural attributes assigned to speakers and speech events, provided that the corpus query tool of choice is able to interpret them.

The nature of the metadata may vary depending on the corpus type and purpose and their public accessibility. Typical EP speaker attributes may be, for example, name and age (when available), gender and function (political affiliation, role in the communicative event), whether or not the speaker is a native speaker of the language of the speech and which regional variety of the language the speaker uses. For speech events, it is common to record duration, topic, mode of delivery (e.g., impromptu, read out, mixed) and speed of delivery. Apart from the aforementioned metadata, EPICG also includes information on the hour of the day when the speech event took place, with the aim to determine how long interpreters had been working up to that point. Finally, metadata on EP interpreters typically include gender and, if it can be determined, the regional variety of the language used by the interpreter. In the case of EPICG, the latter is particularly relevant as the Dutch booth is binational and the language itself is pluricentric with relatively salient differences, including phonetic ones (De Caluwe 2013). 'Comments' are often included with the metadata, allowing the compiler to keep track of one-off features of specific speeches (e.g., that one part of a given speech is inaudible or spoken in a different language) or specifications concerning the speakers' functions.

As EP booths regularly perform relay interpreting when the language spoken is not covered by a given booth (for instance, in the case of a Finnish speaker, the Italian booth might take the relay from the English booth), it makes sense to also specify for any given target speech, whether it is an instance of direct interpretation of the source speaker's speech or an instance of relay interpreting of another booth's output. Relay interpreting is not always easy to detect, but a constant Ear-Voice-Span in excess of 4 s is a fairly reliable predictor. Even more difficult is determining which is the input language, as the only source of information (short of trying to obtain the information directly from the Parliament interpreting services), comes from the sound that is accidently picked up by the interpreter's microphone.

Examples of metadata included with EPIC are provided in Tables 2.2 (speaker's metadata), 2.3 (speech event metadata), and 2.4 (comments).

With regard to interpreters' metadata, EPIC displays the speakers' attributes (by specifying 'speaker: interpreter'), but only the relevant or the publicly known ones are reported (for instance, the attribute 'Political group' is indicated as NA, i.e. not applicable). EPICG, on the other hand, assigns additional specific attributes to interpreters (see Table 2.5).

Interestingly, the values assigned to some speech event-related attributes had to be adjusted to fit the specificity of the material included in EPIC. More specifically, although "duration" and "speech length" were classified as short, medium or long, whereas "speed of delivery" (number of words per minute) as low, medium or high, the actual ranges indicated in Table 2.3 can only be considered valid within the

Table 2.2 Attributes and values assigned to speakers in EPIC, EPICG and EPTIC

ATTRIBUTES	VALUES
Speaker	surname, first name
Gender	F M
Country	Italy …
Mother tongue	Yes No
Political function	MEP MEP Chairman of the session President of the European Parliament Vice-President of the European Parliament European Commission European Council Guest
Political group (according to the verbatim report and EP's website)	Verts/ALE PPE-DE PSE ELDR GUE/NGL UEN TDI EDD NI

context of EP debates, during which 150 w/m can be considered as a "medium" speed of delivery. In EPICG the five points' scale used for duration, text length, and delivery rate is determined on a purely statistical basis. Mode of delivery was assigned depending on whether speakers could be seen reading a script (read mode), speaking without the aid of any written material (impromptu mode), or switching between reading and speaking off-the-cuff (mixed mode).

2.3 Corpus Preparation

In this phase of corpus development the transcribed speeches (and the corresponding verbatim reports and translations, if building an intermodal corpus), together with their metadata, have to be turned into a searchable corpus. Which layers of annotation are added, and how, depends on many factors, such as available resources and skills, research priorities and functionalities of corpus query tools. Here we cover two central ones: part-of-speech (PoS) tagging/lemmatization, and alignment, further subdivided into text alignment and audio-video alignment.

Table 2.3 Attributes and values assigned to speech events in EPIC/EPTIC and EPICG

ATTRIBUTES	VALUES		
	EPIC/EPTIC		EPICG
Duration	Short Medium Long	(<120 s) (120–360 s) (> 360 s)	Very low[a] Low Medium High Very high
Timing	(total number of seconds)		Total number of seconds
Text length	Short Medium Long	(<300 words) (300–1000 words) (> 1000 words)	Very low[a] Low Medium High Very high
Number of words	(total number of words)		Total number of words
Delivery	(number of words per minute)		Number of words per minute
Delivery rate	Slow Medium High	(<130 w/m) (131–160 w/m) (>160 w/m)	Very low[a] Low Medium High Very high
Source text delivery type	Read Impromptu Mixed		Read Impromptu Mixed
Topic (as indicated in the verbatim report)	Agriculture & Fisheries Economics & Finance Employment Environment Health Justice Politics Procedure & Formalities Society & Culture Science & Technology Transport		Agriculture & Fisheries Economics & Finance Employment Environment Health Justice Politics Procedure & Formalities Society & Culture Science & Technology Transport
Specific topic	(as indicated in the verbatim report)		(as indicated on the EP's website)

[a]Very low = values under [mean—(1,5 × SD)]; Low = values between [mean—(1,5 × SD)] and [mean—(0,5 × SD)]; Medium = values between [mean—[(0,5 × SD) + 0,01] and [mean + (0,5 × SD)]; High = values between [mean + [(0,5 × SD) + 0,01] and [mean + (1,5 × SD)]; Very high = values over [mean + (1,5 × SD)]

2.3.1 PoS Tagging and Lemmatization

It is generally agreed that morphosyntactic annotation, i.e. interpretative linguistic information about word classes and base forms (or lemmas) of word tokens is a valuable addition to corpora, "spark[ing] off a whole new range of uses which

Table 2.4 Values of the comment attribute in EPIC/EPTIC

ATTRIBUTE	VALUES
Comment	(specify Council) e.g. Cooperation in the fields of Justice and Home Affairs
	(specify Commission) e.g. Economic and Monetary Affairs
	(specify title) e.g. President of the Republic of Colombia (title: Guest)
	(specify accents) e.g. Irish accent
	technical problems, e.g. 2.53–2.55 (inaudible)

Table 2.5 Attributes and values assigned to interpreters in EPICG

ATTRIBUTES	VALUES
Gender	F M
Booth	NL …
Variety	NL BE
Duration	Total number of seconds
Word count	Total number of words
Delivery	Words per minute
Delivery rate	Very low Low Medium High Very high
EVS	Average EVS during interpretation
Turns	Total number of interpreting turns during session before current interpreting turn and total number of minutes interpreted during session before current interpreting turn

would not have been practicable unless the corpus had been annotated" (Leech 2004, unpaginated).

Depending on the size of the corpus and the resources available in a project, linguistic annotation can be fully automatic, fully manual, or automatic with manual correction. Fully manual POS-tagging and lemmatization are normally impracticable even for small corpus projects, given the time required. At the other extreme, fully automatic taggers/lemmatizers such as the widely used TreeTagger[7] exist for a vast number of languages and provide a relatively straightforward, cost-free solution for enriching a corpus with this type of annotation. However, they are usually credited with accuracy ratings of about 97–98% (Leech 2004), which may or may not be acceptable for one's purposes. Going beyond this performance is likely to require some kind of manual intervention. This may take different forms, e.g. a

[7]http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

subset of automatically annotated texts may be checked manually and then iteratively used to re-train the tagger/lemmatizer, a standard tagset or lemma list may be corrected and fine-tuned to account for the specificities of a given corpus, and rule-based post-processing can be performed on the tagged/lemmatized corpus to sift out residual mistakes.

While linguistic annotation is relatively standard in corpus projects today, it is still an interpretative act. As such, it should not be approached uncritically, particularly when dealing with multilingual spoken mediation data, as is the case with interpreting/intermodal corpora. There are several reasons for this.

First of all, corpus users are likely to expect that linguistic annotation is comparable across source and target text corpora, or across bilingual comparable subcorpora. However, comparable morphosyntactic specifications for different languages are not always available, since most corpus development projects are monolingual. Remarkable exceptions are the SPOOK specifications,[8] providing harmonized tagsets for English, French, German, Italian and Slovene, and the Araneum Universal Tagset (Benko 2016, p. 4247), which provides a comparable "common core" of tags for "11 traditional word classes" in 12 different languages.

Secondly, the expected performance of taggers and lemmatizers on spoken corpora is likely to be much worse than in written corpus projects. Westpfahl and Schmidt (2016, p. 1495) report a POS tagging accuracy of 81.61% when using a standard tagset/parameter file configuration on a corpus of spoken German conversational data. This is because of the lack of typically spoken features in most tagsets and training datasets (e.g. disruptions, unintelligible words, interjections, hesitation markers, onomatopeia).

Finally, though most speeches are delivered and interpreted/translated by native speakers (of the target language), non-native data are also present in EP corpora. Furthermore, it is not unlikely that the mediation process, and interpreting in particular, results in output similar to that found in corpora of non-native (e.g., lingua franca or learner) language (Lanstyák and Heltai 2012). Since non-standard morphosyntactic choices are likely to be found in the corpus, and depending on the aims of the project, decisions may have to be made as to how to deal with differences between form and function. For instance, the choice made by the creators of the VOICE corpus (a spoken English as a lingua franca corpus) was to include both: "for *partly* in the sequence *a partly answer*, we allowed for the tag JJ [adjective], in addition to RB [adverb]" (VOICE 2014, p. 7). Similar solutions may also be relevant for the POS tagging of interpreting corpora.

Whether automatic or manually supervised, linguistic annotation results in two sets of annotations (POS tags and lemmas) accompanying the actual transcripts. In terms of format, it is essential that tags/lemmas are easily separable from the word tokens they refer to, and compatible with the corpus query tool of choice. A widely-used format is the vertical, tab-separated one of the IMS Open Corpus

[8] http://nl.ijs.si/spook/msd/html-en/.

WorkBench (CWB)[9] and related platforms (e.g., the NoSketch Engine (NoSkE), Rychlý 2007), but several alternatives exist, such as XML representations (e.g., < w pos = "JJ" lemma = "real" > real </w >), or the early standard in corpus annotation, that consists in simply including tags and lemmas with the words they refer to, separated by an underscore (e.g., *real_JJ_REAL*). The latter solution has the advantage of being less verbose than other schemes (Leech 2004), and is appropriate for use with simple desktop concordancers like AntConc (Anthony 2014).

2.3.2 Alignment

This Section briefly describes two types of alignment: the process of aligning source and target transcripts with each other and, in the case of intermodal corpora, with the corresponding verbatim reports and (translated) target texts; and alignment of audio/video files with the transcripts. Arguably more straightforward than the process of aligning audio/video files with their transcripts, text alignment can nonetheless be rather demanding, particularly in a multilingual, intermodal project.

2.3.2.1 Text Alignment

A first problem faced by corpus developers is the sheer number of alignments that such corpora require. Taking EPTIC as a case in point, the current version of the corpus features speech transcripts, and the corresponding verbatim reports, in three languages (English translated into French and Italian, and French and Italian translated into English; Bernardini et al. 2016a). Each speech transcript has to be aligned to its interpretation transcript (interpreting subcorpus), and each published verbatim report has to be aligned to its translation (translation subcorpus). Then, to account for the intermodal perspective, each transcript from the interpreting subcorpus has to be aligned to its corresponding written version from the translation subcorpus. The resulting number of alignments approaches two dozen. Unless text alignment is performed fully automatically, with no manual correction (in which case rather poor quality is to be expected, especially when aligning interpretations), the alignment process can take several weeks.

The automatic option has other complications, apart from quality concerns. This is mainly because alignment is typically performed at the level of sentences, and aligners expect sentence boundaries to be present in the texts to be aligned: "[n]owadays, if a bitext is included in a parallel corpus collected for research and/or distribution, we can expect it to be sentence-aligned" (Ahrenberg 2015, p. 398). Yet, sentences are widely acknowledged not to be fully adequate for the segmentation of spoken language (Pietrandrea et al. 2014), especially in interpreting

[9] http://cwb.sourceforge.net/.

corpora such as EPIC and EPICG, where one-to-one correspondence between source and target segments is often missing.

To overcome this obstacle, a compromise can once again be reached between the most appropriate way of representing spoken data in a corpus, and the need of users to easily access source-target and translated-interpreted aligned data. The approach being used in the creation of the trilingual and intermodal EPTIC is described in what follows as an example of such a compromise solution.

In transcribing EPTIC speeches, traditional punctuation marks such as commas and full stops are inserted, along with spoken features such as pauses (...), hesitations and false starts (*time has long bec- beca- ehm arrived.*). This is possible because even the impromptu speeches are rarely fully improvised in the EP setting and generally lack interactional features, making them more akin to written texts than is the case with most other spoken genres.

Alignment of EPTIC files is performed using Intertext Editor (Vondřička 2014), an open source, user-friendly desktop aligner. Intertext Editor relies on Hunalign (Varga et al. 2005) for automatic alignment, but also allows easy manual correction of misalignments and provides several export options (including newline-aligned and TMX). Since multiple alignments are required for the EPTIC corpus setup, the default export format is used, which encodes alignment information as stand-off annotation. Three XML files are produced by Intertext Editor for each bitext alignment: the segmented versions of text 1 and text 2 (source/target or interpreted/translated), and the actual alignment file showing the correspondences between the sentence-like units. Table 2.6 illustrates this with reference to a single aligned unit taken from the English-from-Italian intermodal target sub-corpus.

Table 2.6 EPTIC alignment format produced by Intertext Editor

Interpreted from Italian (text 1)	Translated from Italian (text 2)	Stand-off alignment
<s id = "11" > The confused situation after the flight of President Ben Ali should, or I hope, lead to a situation that we all want to see: social and economic reform which will meet the concerns of the vast majority of the people and the broadening out of the democratic space in the country. </s> <s id = "12" > We need a civil society and a proper democratic pluralist… political system. </s>	<s id = "10" > The new and confused situation that began after the flight of former President, Mr Ben Ali, must now lead to the objective that many have asked for: the start of economic and social reforms to match the expectations of the vast majority of the population and the opening up of democratic forums, to ensure that civil society and the various opposition forces are increasingly involved in public life and in government. </s>	<link type = '1–2' xtargets = '10;11 12' status = 'man'/ >

2.3.2.2 Text-Audio/Video Alignment

The analysis of interpreters' prosody or of Ear-Voice-Span obviously cannot do without audio recordings, as included in the DIRSI Corpus (Bendazzoli 2010) and in EPICG. Including video recordings of source speakers would further allow one to factor in the effects of visual information on the interpreting process.

Several levels of alignment can be used to represent the actual delivery of the speeches. Firstly, each transcript can be aligned with its audio file. This is the minimal option to carry out studies on spoken data, given the fact that transcripts are only a partial representation of the actual data under investigation. The availability of the original audio allows researchers to study prosodic or phonetic features that are impossible to represent in transcripts. Secondly, the analysis of temporal features of interpreting, such as EVS, requires full alignment on three dimensions: source audio/video-target audio/video; source text-target text and audio/video-text. In EXMARaLDA this is achieved by importing bilingual stereo tracks (source language left, target language right), which the system converts into parallel prosograms, as can be seen in Fig. 2.1. The transcription is aligned with the acoustic signal through the creation of "events" (articulated segments, pauses) and event boundaries. Each event corresponds to a segment of the acoustic signal of either source or target audio. Events are allotted a time tag by the system on the basis of the acoustic timeline. Figure 2.1 shows how the different alignments are structured in EXMARaLDA. It shows the start of a speech and its associated interpretation. During the first 4 s of the speech, the interpreter remains silent with the microphone switched off, as can be seen from the identical spectrograms and the aligned transcriptions. At 00.04.1 the interpreter starts rendering the first segment. The time tags just below the centre of the screen are set by selecting portions of the acoustic signal.

The audio-audio and audio-text alignments automatically result in a source text-target text alignment.

Finally, to measure EVS consistently, equivalent lexical items need to be identified at regular intervals. Additional tiers can be created in EXMARaLDA to tag these items, as can be seen at the bottom of Fig. 2.1. EVS can then automatically be extracted by means of a script developed for that purpose.

EXMARaLDA has proved to be a flexible environment to both encode the interpreting data and process its output. Alternative software includes CLAN, ELAN, syncWRITER, TRANSCRIBER, TRANSANA and WINPITCH. Russo et al. (2012) report a detailed description of the pros and cons of two pieces of software tested for the ST-TT/audio/video alignment of EPIC: SPEECHINDEXER (Szakos and Glavitsch 2007), and TRANSANA.[10] Samples of the two different alignment visualizations are shown in Figs. 2.2 and 2.3 (adapted from Russo et al. 2012).

[10] http://www.transana.org/.

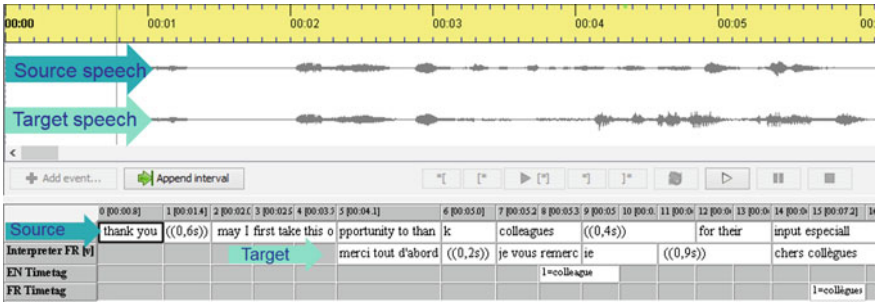

Fig. 2.1 Screenshot from EPICG in EXMARaLDA

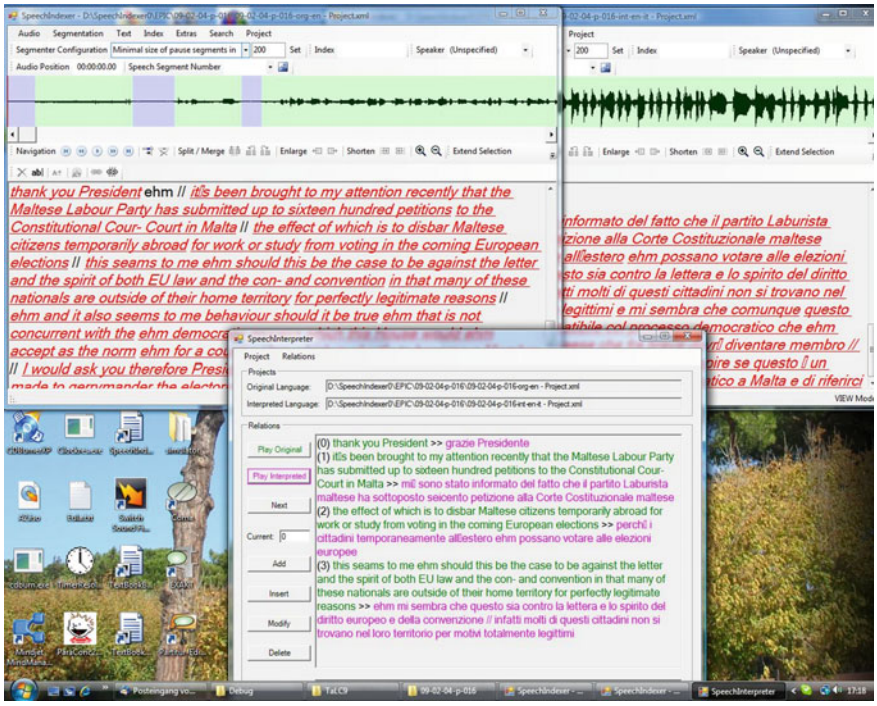

Fig. 2.2 Screenshot of SpeechIndexer

Finally, the minimalist approach that was adopted in EPTIC consists in aligning the sentence-like segments also used for text alignment to their audio/video tracks using subtitling software. Several freely available tools of this kind exist (e.g. Subtitle workshop),[11] that can be adapted to this purpose, since they offer

[11] http://subworkshop.sourceforge.net/.

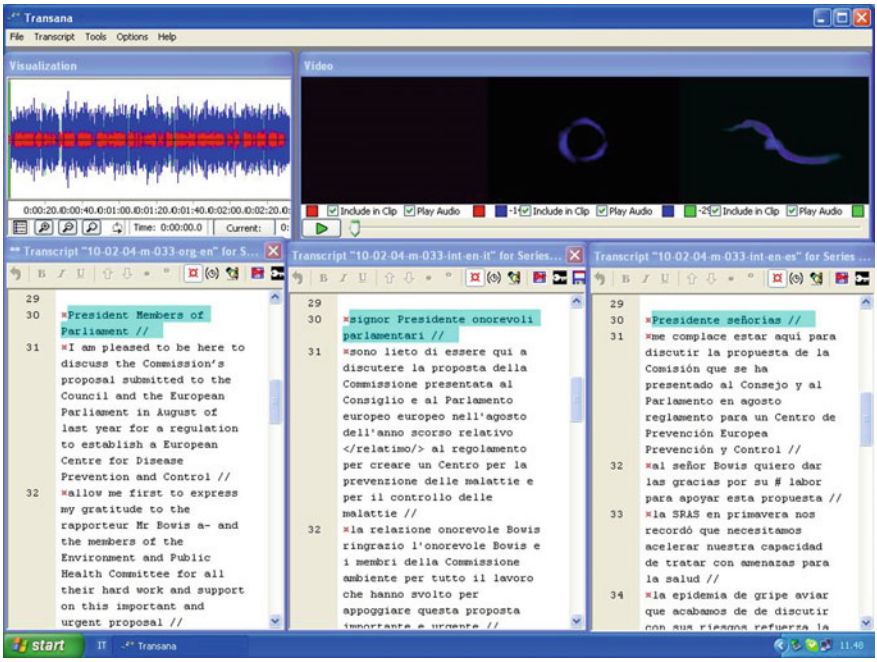

Fig. 2.3 Screenshot of Transana

functionalities to play a video and set start and end times for predefined segments. Start and end times are then easily converted into XML attribute values of the segments themselves.

2.4 The Final Touches: Making the Corpus Ready for Searching

As seen in Sect. 2.3 above, decisions concerning annotation schemes and corresponding file formats should always go hand in hand with the choice of an appropriate query tool for which the corpus will be optimized. Appropriateness should be evaluated in terms of several factors. Some are relevant in all corpus building initiatives, like the usability of the tool by corpus end-users and the familiarity of corpus compilers with complex software architectures. Others are especially crucial for interpreting and intermodal corpora, like the possibility to handle different layers of annotation simultaneously (POS-tagging, lemmatization, and text-audio/video alignment).

Considering only freely available software, several corpus query tools exist, characterized by varying degrees of user-friendliness (both for end-users and corpus compilers), power and flexibility. User-friendly, standalone tools like AntConc or

TextStat[12] are likely to be familiar to most corpus-literate users and can be used out-of-the-box. On the downside, they are not well suited to carry out advanced searches, e.g. searches based on POS patterns or restricted on the basis of contextual metadata, like the speeches delivered by a certain speaker or on certain topics.

At the opposite end of the user-friendliness and power/flexibility spectrum are full-fledged corpus processors like Coma/EXAKT, CWB and the NoSkE. Coma is the corpus compiler provided with EXMARaLDA allowing users to flexibly assemble corpora from existing EXMARaLDA files. Coma comes with a dedicated concordancer called EXAKT. As for CWB and NoSkE, provided that annotation is encoded properly (see Sect. 2.3), these tools make it possible to exploit most of the corpus metadata presented so far to carry out metadata-based queries, and to display (textual) alignments. As is often the case with software, these advanced functionalities come at a cost: performing complex searches usually requires knowledge of a specific search syntax, i.e. the "Corpus Query Language" (Evert et al. 2016) in the case of both CWB and the NoSkE. Moreover, familiarity with Linux/Unix operating systems is necessary on the part of corpus compilers to set up the software infrastructure and to index corpora for use with these tools.

A further strategic feature that is offered by both the CWB and NoSkE environments is the possibility to set up a web-based interface for public consultation of the corpora through a web browser. This allows compilers to maximize uptake of the corpus by the research community, while at the same time maintaining control over accesses to the corpus itself, with no need to distribute its original files (as would be necessary for consultation with standalone tools), and with the option of restricting public access to pre-defined sub-corpora and functionalities.

In what follows, an example is provided of the final format of EPTIC, optimized for indexing and consultation with the NoSkE. While lacking specific features to handle audio and video files, this tool achieves, in our experience, the most favourable trade-off in terms of usability, power and flexibility. As in Sect. 2.3.2.1, the example focuses on EPTIC since it is the most complex of the corpora discussed in this contribution, but the format is applicable to the other members of the EPIC suite, and to similar (EP-based) interpreting and intermodal corpora.

Figure 2.4 presents the format of a text from the EPTIC English-from-Italian target interpreted sub-corpus, showing how the different layers of annotation are encoded in a mix of XML and vertical format. The text header, in XML, contains all the available contextual metadata for the text/speech, the speaker, the source text, and the interpreter; values of attributes for the source text element are only present in the case of interpreted and translated target sub-corpora (and set to "NA" in all other sub-corpora), and the interpreter attribute values are only present in the interpreted target sub-corpora. The body of the text, i.e. in this case the actual transcript, is instead set in the vertical format produced by the TreeTagger, except

[12] http://neon.niederlandistik.fu-berlin.de/textstat/.

Text header

<text id="1003tt-in-en" date="17-01-11-a" length="medium" lengthw="545" duration="medium" durations="232" speed="medium" speedwm="140.8" delivery="interpreted" topic="Politics" topicspec="Statement-by-the-President-of-the-European-Parliament-on-the-situation-in-Tunisia" type="tt-in-en" comments="NA">
<speaker name="Panzeri-Pier-Antonio" gender="M" country="Italy" native="y" politfunc="MEP" politgroup="SD">
<st language="Italian" length="medium" lengthw="648" duration="medium" durations="230" speed="high" speedwm="169.2" delivery="read">
<interpreter gender="M" native="y">

Text body

[…]
<s id="17" timestamp="http://audiovideoserver.org/1003tt-in-en.mp4#t=10.1,13.2">

The	DT	the	the
2008	CD	@card@	2008
initiative	NN	initiative	initiative
is	VBZ	be	is
just	RB	just	just
a	DT	a	a
piece	NN	piece	piece
of	IN	of	of
a	DT	a	a
f-	DYSF	f-	f-
ehm	FPAUSE	ehm	ehm
façade	NN	façade	saçade
with	IN	with	with
no	DT	no	no
real	JJ	real	real
content	NN	content	content

[…]
// SENT //
</s>

Fig. 2.4 EPTIC final format for indexing with the NoSkE

for information on sentences (the "*s*" tag), which is taken from the XML output of the Intertext Editor.[13]

The first three columns in the text body encode, respectively, the normalized text, in which mispronounced words are corrected to their standard forms, the POS of each word, and its lemma; the fourth column contains the non-normalized version of the text (e.g. the case of "façade" in Fig. 2.4). Notice also that three POS tags were added to the TreeTagger default tagset to account for the specificities of the spoken components of EPTIC: DYSF indicates dysfluencies, while FPAUSE and EPAUSE indicate filled and empty pauses respectively.

The "*s*" elements are used not only for text alignment (in particular the "*id*" attribute, see Sect. 2.3.2.1), but also to encode information on the alignment between the transcript and the corresponding video or audio files. Specifically, the "*timestamp*" attribute exploits a convenient feature of the NoSkE that makes it possible to link URLs to each sentence: this feature is used as a workaround to the lack of support for integration of audio and video files into the NoSkE interface. For each sentence in the EPTIC spoken sub-corpora, a URL is provided pointing to an external server (in this case the fictitious "*audiovideoserver.org*"), on which the video file of each speech is uploaded ("*1003tt-in-en.mp4*"); the final part of the URL ("*#t = 10.1,13.2*") specifies the start and end time of the sentence, as determined during the text-audio/video alignment phase (see Sect. 2.3.2.2). When

[13]The outputs of the two tools are integrated (and slightly manipulated) through an ad hoc Perl script, available from the authors upon request.

Fig. 2.5 EPTIC parallel concordances providing access to a time-aligned video within NoSkE

clicking on this link, a browser windows opens and the matching fragment of the audio or video file is played within the browser itself. Figure 2.5 displays an EPTIC sentence as visualized by the concordancer utility of the NoSkE platform.

2.5 Conclusion

In this contribution we have tried to provide as complete an account as possible of the steps involved in designing and building interpreting and intermodal corpora. The family of corpora we have described rely on an accessible data source, the European Parliament plenary debates. These data are in principle easier to collect than most other instances of interpreting, or spoken data in general. Yet the complexity that comes with trying to represent both written and spoken mediation events in an electronic corpus is such that the corpora described here are still too small for many research purposes (e.g., studies of lexis and phraseology), and limited in terms of the number of languages represented (a handful out of the 23 potentially available on the EP website).

We have come clean about the time and expertise required to design the corpus, collect the data, associate them with their metadata (transcribing what needs to be transcribed), annotate them with linguistic tagging, perform multiple text-to-text and text-to-audio/video alignments, integrate the collected/transcribed language samples and the different layers of annotation, and finally make the corpus available through a corpus query tool. And yet we hope that readers are not discouraged from taking up the challenge. The availability of comparable and parallel, spoken and written, mediated and unmediated language data in multiple languages makes the EPIC suite of corpora a unique research object, relevant not only to translation and interpreting scholars, but also to contrastive linguists and, in general, to all those with an interest in usage-based approaches to language studies. A joint effort is needed by a diverse community of corpus linguists to bring the EPIC suite of corpora to maturity. This contribution is meant as a step in that direction.

References

Ahrenberg, Lars. 2015. Alignment. In *Routledge encyclopedia of translation technology*, ed. Sin-Wai Chan, 395–408. London: Routledge.

Angelelli, Claudia. 2012. The sociological turn in translation and interpreting studies. *Translation and Interpreting Studies* 7 (2): 125–128.

Anthony, Laurence. 2014. AntConc (Version 3.4.3). Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/. Accessed 8 June 2017.

Bendazzoli, Claudio. 2010. *Corpora e interpretazione simultanea*. Bologna: Asterisco.

Benko, Vladimír. 2016. Two years of Aranea: Increasing counts and tuning the pipeline. In *Proceedings of the 10th LREC Conference (LREC 2016)*, 4245–4248. Portorož: European Language Resources Association (ELRA).

Bernardini, S., A. Ferraresi, M.A. Lefer, and M. Miličević. 2016a. Simplification in translation and interpreting: Using a tri-directional intermodal corpus to shed light on commonalities and differences. Paper presented at *Translation and Interpreting. Convergence, Contact, Interaction (TransInt).* Trieste, Italy 26–28 May 2016.

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

Davitti, Elena, and Sergio Pasquandrea. 2017. Embodied participation: What multimodal analysis can tell us about interpreter-mediated encounters in pedagogical settings. *Journal of Pragmatics* 107: 105–128.

De Caluwe, Johan. 2013. Nederland en Vlaanderen: (a)symmetrisch pluricentrisme in taal en cultuur. *Internationale Neerlandistiek* 51 (1): 45–59.

Evert, Stefan, and the CWB development team. 2016. The IMS open Corpus Workbench (CWB) CQP query language tutorial version 3.4. http://cwb.sourceforge.net/files/CQP_Tutorial.pdf. Accessed 8 June 2017.

Jefferson, Gail. 2004. Glossary of transcript symbols with an introduction. In *Conversation analysis: Studies from the first generation*, ed. G.H. Lerner, 13–31. Amsterdam: John Benjamins.

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. PhD thesis, Adam Mickiewicz University.

Komter, Martha L. 2006. From talk to text: The interactional construction of a police record. *Research on Language and Social Interaction* 39 (3): 201–228.

Lambert, Sylvie. 1992. Shadowing. *The Interpreters' Newsletter* 4: 15–24.

Lanstyák, István, and Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13 (1): 99–121.

Leech, Geoffrey. 2004. Adding linguistic annotation. In *Developing linguistic corpora: A guide to good practice*, ed. M. Wynne. http://ota.ox.ac.uk/documents/creating/dlc/. Accessed 6 June 2017.

Pietrandrea, Paola, Sylvain Kahane, Anne Lacheret-Dujour, and Frédéric Sabio. 2014. The notion of sentence and other discourse units in corpus annotation. In *Spoken corpora and linguistic studies*, ed. T. Raso, and H. Mello, 331–364. Amsterdam/Philadelphia: John Benjamins.

Plevoets, Koen, and Bart Defrancq. 2016. The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies* 11 (2): 202–224.

Ruhi, Şükriye, Thomas Schmidt, Kai Wörner, and Michael Haugh. 2014. Introduction: Putting practices in spoken corpora into focus. In *Best practices for spoken corpora in linguistic research*, ed. R. Şükriye, M. Haugh, T. Schmidt, and K. Wörner, 1–19. Newcastle: Cambridge Scholars.

Russo, Mariachiara, Claudio Bendazzoli, and Annalisa Sandrelli. 2006. Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: Extended analysis of EPIC. *Forum* 4 (1): 221–254.

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*, ed. F. Straniero Sergio, and C. Falbo, 35–90. Bern: Peter Lang.

Rychlý, Pavel. 2007. Manatee/Bonito: A modular corpus manager. In *Proceedings of the 1st workshop on recent advances in Slavonic Natural Language Processing,* 65–70. Brno: Masaryk University.

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

Shlesinger, Miriam. 2009. Towards a definition of interpretese: An intermodal, corpus-based study. In *Efforts and models in interpreting and translation research: A tribute to Daniel Gile*, ed. G. Hansen, A. Chesterman, and H. Gerzymisch-Arbogast, 237–253. Amsterdam/Philadelphia: John Benjamins.

Sinclair, John M. 1996. EAGLES preliminary recommendations on text typology. www.ilc.cnr.it/EAGLES/texttyp/texttyp.html. Accessed 6 June 2017.

Szakos, Jozsef, and Ulrike Glavitsch 2007. SpeechIndexer in action: Managing endangered Formosan languages. Paper presented at the 8th Annual Conference of the International Speech Communication Association, August 27–31, Antwerp, Belgium.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, 590–596. Borovets.

VOICE. 2014. *Part-of-Speech tagging and lemmatization manual*. Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka.

Vondřička, Pavel. 2014. Aligning parallel texts with InterText. In *Proceedings of the 9th LREC Conference (LREC 2014)*, 1875–1879. Reykyavik: European Language Resources Association (ELRA).

Westpfahl, Swantje, and Thomas Schmidt. 2016. FOLK-Gold—A GOLD standard for Part-of-Speech-Tagging of spoken German. In *Proceedings of the 10th LREC Conference (LREC 2016)*, 1493–1499. Portorož: European Language Resources Association (ELRA).

Wörner, Kai. 2012. Finding the balance between strict defaults and total openness. Collecting and managing metadata for spoken language corpora with the EXMERaLDA Corpus Manager. In *Multilingual corpora and multilingual corpus analysis*, eds. T. Schmidt, and K. Wörner, 383–400. Amsterdam/Philadelphia: John Benjamins.

Zanettin, Federico. 2013. Corpus methods for descriptive Translation Studies. *Procedia—Social and Behavioral Sciences* 95: 20–32.

http://www.springer.com/978-981-10-6198-1

Making Way in Corpus-based Interpreting Studies
Russo, M.; Bendazzoli, C.; Defrancq, B. (Eds.)
2018, XVI, 215 p. 35 illus., 17 illus. in color., Hardcover
ISBN: 978-981-10-6198-1