The Interpreters' Newsletter 2018 (23), 133-149
DOI: 10.13137/2421-714X/22402 ISSN 1591-4127 (print) ISSN 2421-714X (online)
https://www.openstarts.units.it/dspace/handle/10077/2119

The use of comparable corpora in interpreting practice and training

Claudio Fantinuoli
Johannes Gutenberg Universität Mainz/Germersheim
(Germany)

Abstract

Terminology research and domain knowledge acquisition constitute a substantial part of the preparation activity performed daily by professional and trainee interpreters. Corpus-based preparation can assist interpreters in investigating subject-related terminology as well as phraseology and in acquiring subject-specific knowledge. This is particularly important in light of the fact that interpreters often do not have the same level of linguistic and domain expertise as the other event participants. Since tools for corpus analysis have the potential to enhance the quality of preparation, it is reasonable to suggest that they should become an integral part of a modern interpreter's workstation. This paper will introduce two kinds of corpora which can be used in interpreter practice and training in the context of deliberate practice. It will also describe the results of an empirical test of the resources created by a tool designed for this purpose in terms of their adequacy to be used during advance preparation.

Keywords

Interpreter preparation, corpus-based preparation, terminology, computer-assisted interpreting.

Introduction

The use of computer applications is common to all language professions. Although interpreters have been traditionally less prone to accept the introduction of technological advancements (e.g. Tripepi Winteringham 2010; Pym 2011), information and communication technologies (ICTs) have by now found their way in the daily lives of professional and trainee interpreters. From the commissioning of a new assignment to the act of interpreting, the presence of ICTs has become ubiquitous: during advance preparation interpreters search the web for documents related to the topic they are called upon to translate and consult online databases to find translations for unknown terminology (e.g. Tripepi Winteringham 2010); while simultaneously interpreting, they look up glossaries in the booth by means of simultaneous-friendly search tools (e.g. Costa *et al*. 2014; Fantinuoli 2016a); medical interpreters use remote-interpreting platforms to translate patient-doctor interviews (e.g. Nicodemus/Metzger 2014; Andres/Falk 2009); and so forth.

While search engines, electronic dictionaries, and terminological databases find widespread use among novice and experienced interpreters, corpora still seem to be quite unfamiliar to most professionals and trainees. Surveys among professional interpreters have revealed that corpora are ranked quite low among the technological tools and resources used by this professional group (e.g. Corpas Pastor/May Fern 2016). The same consideration applies also to trainees, as it seems that corpora are not widely taught and used in most interpreter training programmes. This is perplexing as over the course of the past decade corpora have been proposed and successfully adopted in providing useful insight into word meaning and use (e.g. Hansen-Schirra *et al*. 2013; Zanettin 2012), in extracting terminology for translation and interpreting tasks (e.g. Fantinuoli 2006) and, if used as a preparation aid, in improving overall terminology rendition during interpretation (e.g. Xu 2015), just to name a few.

In the following sections, it will be argued that more effort should be undertaken to integrate corpora into interpreter training and into professional interpreters' advance preparation. After briefly introducing the role of advance preparation in the interpreting setting (§1), I will look at corpora of political speeches as a tool to support interpreting trainees (§1.1) and at domain-specific corpora for use by professionals to prepare a specialized event (§1.2). Advantages and disadvantages of corpus use will be highlighted. I shall then present how the CorpusMode toolkit[1] can be used to build specialized comparable corpora and to extract terminological information (§2). An evaluation of the proposed method is performed by empirically testing the output of the tool (§3). In Section 4 I will conclude by summarizing my arguments and looking at future prospects.

1 <www.staff.uni-mainz.de/fantinuo/corpusmode.html>.

1. Advance preparation and corpus use

Professionals and trainees usually devote time before an interpreting assignment or class to acquire an overview of the topic and familiarize themselves with the specific terminology and other relevant information (Díaz-Galaz *et al.* 2015). More specifically, during advance preparation, interpreters need to acquire specialized linguistic and domain information in order to bridge the knowledge gap that exists between themselves and the expert speaking, as this has been identified as a prerequisite for succeeding in the task of interpreting (see for example Will (2009) and Fantinuoli (2012)). There is widespread consensus that knowledge acquisition, for example terminology, is better performed when the learning activity is done in context, as new vocabulary can be linked to other words and prior knowledge (Segalowitz/Gatbonton 1995). Against this background, corpora have been suggested to be particularly suitable in supporting the traditional way interpreters prepare for an assignment and in achieving an enhanced interpreting performance (Fantinuoli 2006, 2016b, 2017b; Gorjanc 2009; Xu 2015). They are supposed to be a useful instrument to help interpreters consolidate both the learning of specialized terminology and the acquisition of specialized expertise, the rationale being that *concordance lines* allow learning vocabulary and facts in real context.

The relevance of corpora in advance preparation is not surprising if we take into account the peculiarities of interpretation, the way preparation is traditionally performed, and the resources typically used to accomplish it. Interpreters are called to translate at events dealing with a potentially endless variety of topics, ranging from meetings of quite general nature to highly specialized conferences. Especially in the case of specialized assignments, whether in formal conference settings or face-to-face in hospitals or courts, interpreters often work for specialists who share a linguistic and domain knowledge which is totally or partially unknown to the interpreter. For this reason, as mentioned before, in order to reduce the knowledge gap between the interpreter and the other interlocutors, it is common practice for interpreters to perform preparatory work before the actual act of interpreting (e.g. Gile 2009; Will 2009; Kucharska 2009; Díaz-Galaz *et al.* 2015). This activity is commonly considered a crucial step to guarantee the quality and accuracy which is expected from the interpreter. Although the approach to preparation adopted by interpreters has not been studied extensively and there is only little empirical evidence on how professionals perform this task in their daily work[2], there is a general consensus on the fact that each interpreter has individual preferences and habits about how to prepare an assignment. In general terms, it can be observed that the most common way to prepare for an assignment is by identifying reliable sources of information (texts), extracting relevant information from them, and drawing up a glossary (Díaz-Galaz *et al.* 2015).

The importance of these activities is twofold: interpreters acquire a general idea of the topics to be covered and familiarize themselves with the specific terminology and phraseological items which may be used during the event. Dur-

2 For a discussion of some aspects of this topic, see for example Kalina (2006) and Díaz-Galaz *et al.* (2015).

ing the reading phase, the items whose meanings are unknown or for which a translation in the target language is needed are recorded in a multilingual glossary or are marked in terminological annotations on the preparatory documents (Moser-Mercer 1992; Díaz-Galaz *et al*. 2015). The acquisition of topic-related knowledge alleviates part of the cognitive load during the interpreting phase. A positive effect on quality is expected as it has been hypothesized that this activity facilitates the anticipation and prediction of information. This, in turn, has clear advantages with regard to text comprehension and production as well as the overall interpreting process (Stoll 2009; De Groot 2011; Díaz-Galaz *et al*. 2015).

All of the above mentioned activities are performed nowadays using software tools and on-line resources. As far as linguistic preparation is concerned, especially the establishment of lexical correspondences in several languages, interpreters have a plethora of resources at their disposal, such as online dictionaries, terminology databases and encyclopedias. They provide a wealth of ready-made information, such as definitions, equivalents and examples. One obvious advantage of such resources is that they can be used *out of the box*: they are easy to deploy, quick and provide solutions to many linguistic problems. For these reasons, they can be considered the most common resources used nowadays by interpreters. Despite all the advantages, however, such resources may have at least three disadvantages that need to be mentioned when it comes to meet high quality standards. Firstly, perfect terminology correspondence among two or more languages is difficult to establish. This can be due to different conceptual systems, for example in the case of legal terminology, or because the domains are new and terms have not been coined in all languages (e.g. Will 2009; Xu 2015). Secondly, the high variability and constant evolution of specialized communication make it virtually impossible to find ready-to-use resources available in all possible domains. This means interpreters may not have at disposal any resources for a particular topic, especially for lesser widely spoken languages. Thirdly, available resources often lack contextual information and its "reassuring added value" (Bernardini/Ferraresi 2013: 304) as well as important details such as definitions or examples which are vital to contextualize the information. For the above reasons, interpreters with a limited subject knowledge may not have the means to judge the accuracy or appropriateness of the provided information. As a consequence, its use can potentially lead to sub-optimal performance or even translation errors.

When ready-to-use resources for a specific domain and language are missing, interpreters typically try to establish equivalences between source and target languages, for example at lexical level, by drawing inferences based on term use in real texts. In this case it is quite common to resort to search engines to obtain the frequency of use of terms or expressions (in order to confirm or reject a translation hypothesis) or to search relevant texts and process them manually in order to find solutions to terminology issues. Although this approach may have some advantages, such as the easiness of use of search engines and the fact that new information is mainly processed in context, it also presents several limitations: finding and manually processing adequate texts is quite time consuming, and relying on simple frequencies obtained from search engines allow little to be said about terms because there is no restriction to a specific domain or genre, and so forth.

Some of the limitations of the classical approach to advance preparation can be overcome by integrating the use of corpora, and in particular of ad-hoc comparable corpora. Comparable, ad-hoc corpora are usually collections of similar texts, in most cases of specialized nature, in one or more languages constructed with a specific purpose in mind. Comparability refers to the similarity of the texts being collected, ideally both in terms of topic and text type and genre, to the topic of the assignment, while the ad-hoc designation stresses the fact that they are typically built for a specific interpreting task. Compared to other linguistic resources, comparable corpora offer a number of advantages: they consist of a more comprehensive and diverse variety of source language material and possible translation solutions than dictionaries (Hansen-Schirra/Teich 2009; Zanettin, 2012); they provide samples of real language and are therefore ideal for replicating the jargon used by specific groups of people, such as domain experts; they present lexical items in context with implicit advantages for the learning experience; last but not least, they are a dynamic source of information, as they require an active and intuitive approach to information retrieval. This aligns them perfectly with the cognitive dynamism of the interpreting profession.

Provided the potential usefulness of ad-hoc comparable corpora, there are some potential drawbacks that need to be taken into consideration. Firstly, as they are typically small in size (this does not depend on technical limitations but rather on text availability in specialized domains), ad-hoc corpora may for example not contain matches to help the interpreter verify a working hypothesis. This may discourage the novice user as no immediate and sure rewards can be guaranteed. Secondly, results may be originated by only a few sources and thus may be biased. Thirdly, interpreters generally work under time pressure. In their eyes embarking on a corpus analyses activity may appear too laborious, as it requires an investment of time to engage actively in the discovery mechanism typical of corpus analysis. Furthermore, corpora typically need to be created ad-hoc by the interpreter as they need to be tailored to the topic. This activity must be performed with systems designed specifically for this task. This may be a deterrent for many interpreters, as they need to embark upon learning a new (and often complex) tool.

In the following two sections, two types of comparable corpora are introduced: speech corpora for use in interpreter training (§1.1) and domain-specific corpora for use in a professional context (§1.2).

1.1 Speech corpora in interpreter training

One area in which corpora can be successfully deployed is interpreter training. In the last few years much attention has been devoted to the methods for improving student skills and expertise. Becoming a professional interpreter involves the mastery of not only the interpreting process, but also languages, cultures, specific domain knowledge, and so forth. The development of such expertise, given the complex nature of interpreting, requires extensive experience, dedication, and practice. However, merely practising a skill repetitively does not invariably lead to expert performance, which is the ultimate goal of conference interpreting train-

ing programmes. As a means to improve performance and eventually lead to desired proficiency, Ericsson (2006: 692) proposes *deliberate practice* which he defines as "tasks that are initially outside of their current realm of reliable performance, yet can be mastered within hours of practice by concentrating on critical aspects and by gradually refining performance through repetitions after feedback" . The acquisition of competence, especially if the skills are of a complex nature, seems to be more efficient when the task is broken down into smaller units that can be addressed separately during training. In deliberate practice, sub-skills are identified within a broader skill that need to be developed and are addressed explicitly by trainers and trainees before they are combined in what is the final performance.

One of these units is the set of language-related competences trainees need to acquire with the goal of improving their ability to produce adequate renditions of a given original text (among others the expansion of semantic groups, use of adequate registry, etc.). Typically, such competences are analyzed and trained during interpreting classes and in particular during the feedback session in which students' performances are assessed by the teacher. There is no doubt that this approach still represents a good didactic practice. However, trainees often lack instruments to take increased responsibility for their own learning rather than being taught in a more passive mode. To overcome this, such learning activities can be integrated by a corpus-driven approach. The use of corpora in and outside the interpreting classroom is in line with a constructivist approach to learning. In this approach knowledge is constructed by learners, rather than simply transmitted to them by teachers. As seen above, in the interpreting classroom it is the teacher who traditionally is the main authority, both in terms of performance evaluation and problem solving. Constructivism proposes a shift of authority, giving the student more responsibility, autonomy, and control in the learning process. Students are encouraged to engage in self-reflection about their skills and to develop metacognitive abilities to monitor and evaluate their performance, for instance on the language level. In this context, corpora can be the ideal instrument in allowing trainees to be at the centre of the (language and domain-related) learning process. They can be a source of potentially endless linguistic experience. Embarking on a corpus analysis exercise in order to find answers to real translation problems and to assess the linguistic component of an interpreting performance can improve both comprehension as well as rendition, on one side, and encourage trainee motivation to excel, on the other.

As seen in Section 1, through corpora students can embark on a rewarding process of discovery and exploration. For example, when assessing their own performance and identifying their weaknesses, they are able to verify autonomously the linguistic choices they have made, both at lexical, syntactical and grammatical level. Furthermore, this discovery process favours the acquisition of specific knowledge about the topic, as words and expressions are always seen embedded in their semantic context. Thanks to the a serendipity process (Johns 1988), as one term can lead to another depending on the user's intuition and needs, students can easily extend their knowledge of vocabulary, find synonyms, alternative phraseology, etc., i.e. they are able to increase their linguistic flexibility. When learning to interpret into their foreign language, for example, students may verify not only translation hypotheses and grammar structure, but also idi-

omaticity. This appears quite important because of the (linguistic) difficulties of interpreting into a foreign language pointed out in the literature (Seleskovitch/Lederer 1989).

In the past, several corpora types, for example parallel[3] or interpreting[4] corpora, have been envisaged to be useful in interpreter training (cf. Sandrelli 2010). Considering the type of texts typically used in conference interpreter training and the language-related abilities that need to be acquired by trainees, another special genre of corpus can be considered particularly useful while learning interpretation: comparable corpora of original speeches[5]. Such corpora are collections of the transcriptions of speeches delivered by politicians, scientists, etc. in formal contexts. There are several advantages of the use of this kind of corpora over the ones generally proposed in the literature (cf. Sandrelli 2010). Firstly, the fact that they contain the same genre of textual material as the texts students are called upon to interpret makes the identification of linguistic features which may be of interest for the student more easy, as they are frequent and, therefore, easier to spot. Secondly, the use of original, non-translated texts (as in the case of parallel corpora or interpreting corpora) gives students the opportunity to work on material which has not been influenced by the translation process. This facilitates the use of an unbiased, natural language. Finally, trainees are likely to benefit from the exposure to original texts uttered by experts, politicians, etc. in the target language because they provide a model to support the production of a natural target language speech.

Speech corpora can be used as a tool to support deliberate practice and autonomous learning, helping students deepen language and discourse competences in the classroom. The outcome of a learning experience based on corpora can be extremely rewarding, both in terms of interpretation-related language acquisition as well as in the emancipation of the trainee from the central role played by the trainer, as far as the appropriateness of linguistic choices is concerned. In fact, the student, if properly accompanied in this emancipation process, can improve his or her ability to self-assess linguistic performance and improve it.

1.2 Exploiting domain-specific corpora

For professional interpreters, particularly useful types of corpora are comparable, domain-specific corpora, i.e. collections of texts which represent and reflect the language of a particular topic. They normally include texts dealing with one specific subject (e.g. bio-energy), while other parameters, such as genre (e.g. handbook or leaflet), size or language variety, may vary according to the objective and the design criteria (Zanettin 2012). To some extent, a domain-specific corpus

3 Parallel corpora contain original speeches and their respective translation.

4 Interpreting corpora contain speeches interpreted by interpreters who are (generally) native speakers.

5 Open access monolingual corpora of political speeches for several European languages are freely available at <www.staff.uni-mainz.de/fantinuo/speechcorpora>.

is similar to the domain-specific collection of texts that interpreters usually use as reference material during preparatory stages of an assignment with the goal of gaining insight into a subject and finding relevant terminology and phraseology. The main difference is the quantity of the texts and the way they can be consulted: digital or paper documents may be consulted in a linear way, i.e. read partially or entirely as a normal text (by means of word processing search facilites, in the case of digital documents). The consultation of texts on the web, in order to draw inferences based on actual use in context, for example, is a time consuming and exacting process, which requires rapid evaluation of the reliability of sources, the opening and closing of multiple pages, and an acceptance of the many limitations of search engines that were not designed for linguists. Texts contained in a corpus, however, can also be consulted in a non-linear way: thanks to the query and display feature of corpus analysis tools, the user can approach the text in a *bottom-up* manner, starting from the terminology/phraseology of the domain to the creation of its conceptual structure. Furthermore, most tools allow full-text browsing and offer the rich contextual information required for decision-making (Bernardini/Ferraresi 2013). This allows interpreters to explore the textual material of a specialized subject in a dynamic, interactive and explorative way (Fantinuoli 2006, 2017b) as they no longer need to browse through different texts and pages.

When discussing the use of domain-specific comparable corpora in translation, many scholars have pointed out their practical advantages. Thanks to the repetitiveness of specialized texts, for example in terms of lexical items, domain-specific corpora are ideal for automatizing the search for patterns in texts and, therefore, for finding useful information (because very frequent). Aston (2000), for instance, stressed how disposable domain-specific corpora facilitate data interpretation as the likelihood of finding ambiguous data is small. Furthermore, the bottom-up search strategy implied in corpus use – from the terminology to the conceptual structure of a domain – improves the way hypotheses are formulated and validated (e.g. Fantinuoli 2006). Varantola (2003) stresses the reassurance role of corpora, as the process of decision-making (for example, choosing one translation among several alternatives) is supported by empirical evidence. It is reasonable to suggest that all the advantages indicated for translators may be extended to interpreters during the preparatory stage of an assignment.

The discovery experience is based on the serendipity process introduced above. Thanks to corpora, interpreters may become the active subject of a data-driven learning process[6] which offers "virtually unlimited opportunities for learning by discovery, as learners embark on challenging journeys whose outcomes are unpredictable and usually rewarding" (Bernardini 2001: 246). For interpreters, the reward will be the ability to construct conceptual structures in the field of interest, discovering concepts and terms related to each other, and consequently being able to construct – before the event – an internal representation of the particular subject, with obvious advantages for the successful outcome of

6 For more information on the Data Driven Learning approach, see for example Boulton (2009).

the interpreting task, and, on a more practical level, a terminology list to be used during the course of the event.

2. Building ad-hoc corpora and extraction of linguistic information

As argued in Section 1, interpreters' needs in terms of linguistic and extra-linguistic preparation are quite specific. Thus, in order to address those needs with a corpus-based approach, adequate resources must be made available. Only very rarely are domain-specific corpora available for immediate use. Hence, interpreters wanting to use corpora in the course of advance preparation need to create their own resources each time they engage with a new assignment, a new subject, or a new client. If professional interpreters and interpreting trainees are supposed to successfully employ corpora, the time and the effort required to create ad-hoc corpora need to be as little as possible. Only if the trade-off between the resources invested and the output is perceived as positive, will interpreters engage in a corpus-oriented approach.

There are several tools that allow for a (semi)automatic collection of specialized texts from the web, including BootCat[7], AntCorGer[8] and SketchEngine[9]. They are very powerful programmes, however, they have not been developed with the interpreter in mind. As a consequence, they may appear cumbersome or limited in the functionalities they offer to the interpreter. In this section, I propose to deploy CorpusMode (Fantinuoli 2017b), a free and interpreter-oriented programme that automates the process of finding reference texts, extracting information from a corpus and analyzing it. The tool makes use of the Cognitive Services offered by Microsoft, in particular Bing Web Search Api[10], to find domain-relevant documents on the web. It downloads and transforms the texts in an ad-hoc built corpus, extracts specialized terminology and collocations, and allows the user to explore the corpus through a dynamic concordancer.

The basic procedure of corpus construction implemented in CorpusMode is straightforward and starts from a small list of single- or multi-word terms (called *seeds*) that are expected to be typical of the domain of interest. Appropriately combined, the seeds are used as a query string to search the web for relevant documents. To prevent the collection of unrelated texts, the seeds provided by the user should ideally be unambiguous, highly specialized and typical only in the domain of interest. To influence the characteristics of the corpus building procedure, the user can specify a set of further search parameters, such as domain (.com, .de, europa.eu and so forth), language, format (PDF and/or HTML) and number of texts to be downloaded. A list of URLs is retrieved and presented to the user for manual assessment. In this stage the user can simply decide to retain or delete a text candidate by assessing the URL itself or inspecting the documents related to it. Finally, the accepted URLs returned for each query are

7 <http://bootcat.dipintra.it>.
8 <http://www.laurenceanthony.net/software/antcorgen/>.
9 <www.sketchengine.co.uk>.
10 <www.microsoft.com/cognitive-services/en-us/bing-web-search-api>.

retrieved, cleaned, and imported into a corpus. The dimension of the corpus collected this way depends on many factors, such as the number of texts selected by the user, the number of documents automatically discarded due to copyright protection, and the quantity of texts contained in the single documents.

Informal tests have proven that it is possible to build large domain-specific corpora containing several hundreds of texts by use of default parameters and without manual intervention[11]. To build a corpus of around 80-100 texts, which can be considered a standard size for a domain-specific corpus, the process typically requires only a few minutes (but time depends on many factors, such as the speed of the Internet connection, the server, file dimensions etc.). Although the quality of such "quick-and-dirty" corpora[12] may be variable and cannot be compared to the typical quality of supervised, well-constructed and balanced corpora, the proposed approach has been indicated to be most suitable for the creation of corpora for interpreters (Gorjanc 2009; Fantinuoli 2006) and has already been successfully used in both translation and interpreting settings (Xu 2015; Castagnoli 2006).

In a corpus-based preparation approach, interpreters need specialized terminology both to create multilingual glossaries as well as to perform the corpus browsing activity which is at the basis of the serendipity process described in Section 2. In fact, this discovery activity can be largely improved if terminology-driven (e.g. Xu 2015; Fantinuoli 2017a). A first list of terms can be obtained by means of automatic terminology extraction. This type of extraction from a corpus poses several challenges. CorpusMode implements a hybrid extraction method which combines linguistic knowledge and statistical measures[13]. To improve the usability of the software for interpreters, the focus of the terminology extraction algorithm is on precision rather than on recall[14]. Accordingly, in order to be more suitable for interpreters (who require immediately usable results), the tool was designed to reduce the number of ill-formed constructions, even at the risk of missing eligible candidates, and to retain only highly specialized and frequent terms. The term selection principle is based on the assumption that single-word and multi-word terms have a certain fixed set of linguistic structures, for example "Noun + Preposition + Noun". The tool assigns a part-of-speech (POS) tag to each word and extracts all candidate terms that adhere to predefined patterns. The resulting term list is then filtered by means of statistical measures (e.g. relative frequency in the constructed specialized corpus vs. relative frequency in a general reference corpus to determine if a term is specific for the domain under analysis), and heuristics (e.g. common vocabulary can be excluded from the final list), in order to rank the candidate terms and select the most appropriate. The final list of terms can be adapted on the basis of the interpreters' profile.

11 See also Bernardini/Ferraresi (2013) for similar observations on tests conducted with BootCat.

12 In Tribble's sense, "quick-and-dirty" are corpora informally produced for an immediate use without elaborate deliberations about their compositions (Tribble 1997).

13 For a detailed description of the automatic extraction procedure, see Fantinuoli (2012).

14 Precision is the ratio between correct and incorrect terms retrieved in the corpus. Recall is the ratio between the term actually extracted and all terms that ideally should have been extracted.

For example, a novice interpreter or an interpreter who has never worked in a particular domain, may wish to include both specific terminology of the domain and general terms that are frequently used in a particular field, while an expert interpreter, may retain only the very specialized terminology of the subject he/she is called upon to interpret (cf. Fantinuoli 2017a).

Besides terminology, a list of collocations can be automatically generated for any given term. The function aims at identifying the most frequent collocates of a term inside a corpus (and thus in the specific domain), leaving out less typical collocational patterns. This is in line with the assumption that interpreters usually need to manage the most typical and, therefore, most frequent linguistic information for a given term. Collocations are identified by applying frequency measures of the POS pattern of interest which occur in the span of a defined window of terms. The most frequent results are presented to the user as a list of collocations and of their frequency in the corpus.

An empirical evaluation of the corpus building and the terminology extraction procedure is discussed in the next section.

3. Evaluation

In this section, CorpusMode is evaluated in terms of the quality of the constructed corpus and of the extracted terminology by applying it to a specific domain (biogas) for the English and German language combination. The proposed empirical test does not claim to be exhaustive. As noted above, the variables at stake are many and the outcome depends on many factors, such as the domain, the availability and quality of documents in a specific language, the selection of the initial seeds, the number of documents retrieved, the ranking list of web pages retrieved by the Microsoft Services at the moment of testing, among others. This makes it hard to create an empirical setting which allows for easy replication. However, this preliminary evaluation can provide some general indications about the advantages and limitations of the tool and the procedure proposed.

With the corpus building method described in Section 2 two, comparable monolingual corpora for the topic *biogas* were automatically constructed. The topic was selected randomly out of a list of 20 possible subjects. For the two corpora (in English and German), two lists of specialized terms were automatically extracted using the algorithm presented in Section 2. In both tasks the default settings were used.

The seeds needed for the corpus construction were obtained by selecting two specialized terms in the Wikipedia entry for *biogas*. This task is meant to replicate the method an interpreter might use to obtain the keywords for the corpus building routine[15]. The seeds were combined into two different word combinations made up of the topic descriptor (Biogas) and the selected terms, as indicated in Table 1. The default search parameters were used (file type: PDF, domain: not

15 In the case of simultaneous interpreting, Fantinuoli (2017b) proposed to select the seeds from the titles of the events the interpreters are called to interpret.

specified, number of URLs: max). No manual filtering of results was performed. The size of the two corpora is provided in Table 2.

English	German
biogas + anaerobic digestion	Biogas + Vergärung
biogas + renewable energy	Biogas + Erneuerbare Energie

Table 1. Seeds used for corpus construction

	English	German
N. of URLs automatically collected	100	96
N. of texts successfully downloaded	99	94
N. of texts successfully converted	88	79
N. of tokens in final corpus	951,023	722,896

Table 2. Size of the corpora

The evaluation methodology applied to assess the output of the corpus building procedure is similar to the one described in Bernardini/Ferraresi (2013), with 30 students of interpretation (both graduate as well as undergraduate) acting as informants. They were asked to evaluate a randomly extracted list of 10 texts for each corpus (which corresponds to 11.36% and 12.66% of the total URLs successfully downloaded, converted and inserted in the corpus). The selected texts were presented to the informants in a random order with the request to briefly read the beginnings of the texts and rate them according to their relatedness to the topic and suitability as reference and preparatory material. This kind of empirical test aims at evaluating the perceived suitability of the texts[16]. There were four possible answers (definitely yes, probably yes, probably no, definitely no). For a text to be considered as relevant, more than 50% of the informants had to rate it definitely or probably appropriate. As Table 3 shows, almost all texts passed this evaluation stage.

	English	German
Relevant texts	9	10

Table 3. Number of relevant texts in the two corpora

The results of the corpus building procedure are quite encouraging in terms of output quality. All German texts and almost all English ones were ranked as related to the topic and useful for advance preparation. If the texts are not considered singularly but as a corpus, the percentage of appropriate texts included in the corpus reaches 84% for English and 86% for German. Our analysis thus confirms

16 In order to test not only the perceived suitability of the collected texts but also their actual usefulness, it would be necessary to back this up with an experiment in which 50% of the subjects would prepare for an assignment by using "traditional" methods and 50% would do it via a do-it-yourself corpus (see Xu 2015).

the results of prior studies (Fantinuoli 2006; Xu 2015) according to which the automated construction of domain-specific corpora can help interpreters to quickly obtain specialized documents for use in corpus-based advance preparation.

The evaluation of the term extraction algorithm is based on a categorization system similar to the one proposed by Fantinuoli (2006). In particular, the extracted terms were divided into three groups according to their level of specialization and well-formedness: 1) specialized terms; 2) general terms; and 3) incomplete or ill-formed words. The top 50 terms extracted by the tool using the default settings were evaluated accordingly. Table 4 gives a summary of the results.

	English	German
1) Specialized terms	23	13
2) General terms	16	25
3) Ill-formed terms	11	12
Total	50	50

Table 4. Evaluation of the extracted terms

The evaluation of the terminology obtained with the tool is not completely satisfactory. The percentage of specialized terms extracted in the experiment is quite low and as a consequence the final list of candidate terms appears biased towards general or ill-formed terms. In particular, the high number of ill-formed terms shows that the identification of relevant terminology is still a problem, at least for the tool used, if the corpus is "quick-and-dirty", such as the one constructed during this test. Informal observations of the corpus data, in fact, show that the collected and prepared texts presented a lot of 'noise' (poor formatting, incorrect hyphenations, improper language, etc.). This has a negative impact on the process of term extraction and, as a consequence, on the quality of the outcome. The results highlight also another possible drawback connected with the high percentage of general words extracted. Although general words, especially the high frequency ones, may be useful to many interpreters (Xu 2015), it is quite difficult to automatically distinguish useful general terms (for example terms which are highly frequent in the domain of interest even if they are not specialized, such as 'renewable resource'), from frequent terms that are less interesting for the interpreting process (such as 'energy'). Further empirical tests using a more detailed annotation system, such as the one proposed by Xu/Sharoff (2014), could be used to better understand the needs of interpreters in terms of general, non-specialized words. In this experiment, the evaluation of the collocation extraction has been left out from the test since candidate collocations can be extracted for any term of interest, adding another layer of subjectivity to the experimental setting.

4. Conclusions

In this paper an interpreter's perspective on the use of comparable corpora has been presented. It has been argued that the use of comparable corpora can be integrated in the interpreting workflow to extend the set of resources typically

used by interpreters, such as electronic dictionaries and terminology databases. Furthermore, it has been suggested that comparable corpora may be particularly suitable for the interpreter preparation. Two kinds of corpora were presented: ad-hoc specialized corpora, which can be used in a professional setting to acquire linguistic and specialized domain knowledge, and corpora of speeches, which are suitable in interpreter training to support deliberate practice and help students in gaining the language expertise required to perform well at the end of their university programme.

The lack of ready-to-use, open access corpora constitutes a limit to the deployment of corpus-based techniques in both interpreting practice and training. For this reason, a software allowing an almost effortless way to construct specialized corpora and a tool to automatically extract relevant terminology have been presented. An empirical test has been conducted to evaluate if the texts collected by the tool are perceived as useful for interpreting tasks. The extracted terminology has been evaluated according to its level of specialization and well-formedness. The relevance of the retrieved texts seems suitable to satisfy the needs of interpreters, though some limitations have been found in the quality of the terms extracted. One possible reason is that the collected texts were too 'noisy' to allow for clean processing during the terminology extraction procedure.

While we have suggested that comparable corpora should be seen as complementary resources to be used in different ways and for different purposes during the preparation stage, there is no doubt that in order to gain the favour of interpreters a corpus-based approach requires tools perfectly tuned to interpreters' requirements, especially in terms of easiness of use, speed and flexibility. In order to make the corpus construction procedure more intuitive, the possibility of further simplifying the process should be investigated. For example, instead of selecting appropriate seeds to start the corpus building routine, users could define the domain of the corpus to be collected by simply indicating a relevant Wikipedia entry or a particular website. The selection of the seeds needed to find the relevant URL and the corpus building procedure will be performed automatically without human supervision. To increase the usefulness of the terminology extraction feature proposed, which depends highly on the 'cleanness' of the corpus, there is need to explore new ways of avoiding the incorporation of texts containing poor formatting or multiple languages. Finally, the possibility to implement a term alignment method for the monolingual term lists extracted from the two comparable corpora should be pursued as a means to obtain the first draft of a bilingual glossary.

Hopefully, some of these improvements will raise interest in the proposed approach for the interpreter community.

References

Andres D. / Falk S. (2009) "Information and communication technologies (ICT) in interpreting – remote and telephone interpreting", in D. Andres / S. Pöllabauer (eds) *Spürst Du wie der Bauch rauf runter? / Is Everything all Topsy Turvy in your Tummy? - Fachdolmetschen im Gesundheitsbereich / Health Care Interpreting*, München, Martin Meidenbauer, 9-27.

Aston G. (2000) "I corpora nella didattica della traduzione", in S. Bernardini / F. Zanettin (eds) *I corpora nella didattica della traduzione*, Bologna, CLUEB, 21-29.

Bernardini S. (2001) "Spoilt for choice: a learner explores general language", in G. Aston (ed.) *Learning with Corpora*, Bologna, CLUEB, 220-249.

Bernardini S. / Ferraresi A. (2013) "Old needs, new solutions: comparable corpora for language professionals", in S. Sharoff / R. Rapp / P. Zweigenbaum / P. Fung (eds) *Building and Using Comparable Corpora*, Heidelberg, Springer, 303-319.

Boulton A. (2009) "Data-driven learning: reasonable fears and rational reassurance", *Indian Journal of Applied Linguistics* 35/1, 81-106.

Castagnoli S. (2006) "Using the Web as a source of LSP corpora in the terminology classroom", in M. Baroni / S. Bernardini (eds) *Wacky! Working Papers on the Web as Corpus*, Bologna, GEDIT, 159-172.

Corpas Pastor G. / May Fern L. (2016) *A Survey of Interpreters' Needs and their Practices Related to Language Technology. Technical report*, Universidad de Málaga.

Costa H. / Corpas Pastor G. / Durán Muñoz I. (2014) "A comparative user evaluation of terminology management tools for interpreters", in *Proceedings of the Workshop on Computational Terminology (CompuTerm'14)*, 68-76.

Díaz-Galaz S. / Padilla P. / Bajo M. T. (2015) "The role of advance preparation in simultaneous interpreting: a comparison of professional interpreters and interpreting students", *Interpreting* 17/1, 1-25.

De Groot A. M. (2011) *Language and Cognition in Bilinguals and Multilinguals: An Introduction*, New York, Psychology Press.

Ericsson K. A. (2006) "The influence of experience and deliberate practice on the development of superior expert performance", in K. A. Ericsson / N. Charness / P. J. Feltovich / R. R. Hoffman (eds) *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge, Cambridge University Press, 683-704.

Fantinuoli C. (2006) "Specialized corpora from the Web for simultaneous interpreters", in M. Baroni / S. Bernardini (eds) *Wacky! Working papers on the Web as Corpus*, Bologna, GEDIT, 173-190.

Fantinuoli C. (2012) *InterpretBank - Design and Implementation of a Terminology and Knowledge Management Software for Conference Interpreters*, unpublished PhD thesis, University of Mainz.

Fantinuoli C. (2016a) "InterpretBank. Redefining computer-assisted interpreting tools", in *Proceedings of the Translating and the Computer 38 Conference*, London, Editions Tradulex, 42-52.

Fantinuoli C. (2016b) "Revisiting corpus creation and analysis tools for translation tasks", *Cadernos de Tradução*, 36/1, 62-87.

Fantinuoli C. (2017a) "Computer-assisted preparation in conference interpreting", *Translation and Interpreting*, 9/2, 24-37.
Fantinuoli C. (2017b) "Computerlinguistik in der Dolmetschpraxis unter besonderer Berücksichtigung der Korpusanalyse", in S. Hansen-Schirra / S. Neumann / O. Čulo (eds), *Annotation, Exploitation and Evaluation of Parallel Corpora*, Berlin, Language Science Press, Berlin, 111-146.
Gile D. (2009) *Basic Concepts and Models for Interpreter and Translator Training: Revised edition*, Amsterdam/Philadelphia, John Benjamins Publishing Company.
Gorjanc V. (2009) "Terminology resources and terminological data management for medical interpreters", in D. Andres / S. Pöllabauer (eds) *Spürst Du, wie der Bauch rauf-runter? Fachdolmetschen im Gesundheitsbereich. Is Everything all Topsy Turvy in your tummy? Healthcare Interpreting*, München, Meidenbauer, 85-95.
Hansen-Schirra S. / Neumann S. / Steiner E. (eds) (2013) *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*, Berlin, de Gruyter.
Hansen-Schirra S. / Teich E. (2009) "Corpora in Human Translation", in A. Lüdeling / M. Kytö (eds) *Corpus Linguistics - An International Handbook*, Berlin, de Gruyter, 1159-1175.
Johns T. (1988) "Whence and whither classroom concordancing", in T. Bongaerts / P. de Haan / S. Lobbe / H. Wekker (eds) *Computer Applications in Language Learning*, Dordrecht, Foris, 9-27.
Kalina S. (2006) "Zur Dokumentation von Maßnahmen der Qualitätssicherung beim Konferenzdolmetschen", in C. Heine / K. Schubert / H. Gerzymisch-Arbogast (eds) *Translation Theory and Methodology*, Tübingen, Gunter Narr Verlag, 253-268.
Kucharska A. (2009) *Simultandolmetschen in defizitären Situationen. Strategien der translatorischen Optimierung*, Leipzig, Frank & Timme.
Moser-Mercer B. (1992) "Terminology documentation in conference interpretation", *Terminologie et traduction*, 2/3, Office des publications des Communautés européennes.
Nicodemus B. / Metzger M. (2014) *Investigations in Healthcare Interpreting*, Washington D.C., Gallaudet University Press.
Pym A. (2011) "What technology does to translating", *Translation & Interpreting*, 3/1, 1-9.
Sandrelli S. (2010) "Corpus-based interpreting studies and interpreter training: a modest proposal", in L.N. Zybatow (ed.) *Translationswissenschaft: Stand und Perspektiven*, Frankfurt, Peter Lang, 69-90.
Segalowitz N. / Gatbonton E. (1995) "Automaticity and lexical skills in second language fluency: implications for computer assisted language learning", *Computer Assisted Language Learning*, 8/2-3, 129-149.
Seleskovitch D. / Lederer M.L. (1989) *A Systematic Approach to Teaching Interpretation*, Luxembourg, Didier Erudition.
Stoll C. (2009) *Jenseits simultanfähiger Terminologiesysteme*, Trier, Wvt Wissenschaftlicher Verlag.

Tribble C. (1997) "Improvising corpora for ELT: quick and dirty ways of developing corpora for language teaching", Lodz, Lodz University Press, 106-117.
Tripepi Winteringham S. (2010) "The usefulness of ICTs in interpreting practice", *The Interpreters' Newsletter*, 15, 87-99.
Varantola K. (2003) "Translators and Disposable Corpora", in F. Zanettin, / S. Bernardini / D. Stewart (eds) *Corpora in Translator Education*, Manchester, St. Jerome Publishing, 55-70.
Will M. (2009) *Dolmetschorientierte Terminologiearbeit. Modell und Methode*, Tübingen, Gunter Narr Verlag.
Xu R. (2015) *Terminology Preparation for Simultaneous Interpreters*, unpublished PhD thesis, University of Leeds.
Xu R. / Sharoff S. (2014) "Evaluating term extraction methods for interpreters", in *Proceedings of the 4th International Workshop on Computational Terminology*, Dublin, 86-93.
Zanettin F. (2012) *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*, Manchester, St. Jerome Publishing.