"Corpus-based Interpreting Studies: a booming research field", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2304

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy. This printout was generated directly from the online version of this article and can be freely distributed under Creative Commons License **CC BY-NC-ND 4.0**.

Corpus-based Interpreting Studies: a booming research field

By Claudio Bendazzoli, Mariachiara Russo & Bart Defrancq (University of Turin, University of Bologna, & Ghent University)

Abstract & Keywords

Keywords:

Corpus-based interpreting research has gained considerable momentum over the last few years. Indeed, an increasing number of scholars have developed corpora using data from different settings or taken advantage of existing ones. After refining the methodology to address the many challenges involved in the corpus-based approach, investigations carried out within this research paradigm are providing insightful observations about the interpreting process and product, including comparisons between different Translation modes, that is oral interpretation and written translation. In addition, corpora are now being developed and used as educational resources, thus giving trainee interpreters access to principled sets of materials for targeted practice as well as opportunities to reflect upon the skills they are acquiring.

This special issue presents novel investigations that are pushing corpus-based interpreting research to the next level. Some of these are based on, or are deeply inspired by, existing Corpus-based Interpreting Studies (CIS) projects, such as the pioneering European Parliament Interpreting Corpus (EPIC), while others endeavor to embrace other types of interpreting from more sensitive communicative settings, such as health care and court interpreting.

The aim of this special issue is to provide a forum to share the results obtained and the efforts being made in a booming research field, which, as editors, we believe deserves even further support and dissemination.

The seven contributions included here are organized into two main sections relating to two complementary research areas: interpreting practice and interpreter training.

Section 1 *Interpreting practice: developing and using corpora to study interpreting* includes papers focusing on simultaneous interpreting, with the exception of the essay by Sara Castagnoli and Natacha Niemants "Corpora worth creating: A pilot study on telephone interpreting". Notwithstanding the many limitations entailed in the creation of this small corpus, the two authors experiment with different types of annotation and lend support to the usefulness of applying the corpus-based approach even to limited collections of data. Turning to simultaneous interpreting, three papers are based on European Parliament (EP) simultaneous interpreting data. In her study "The translation challenges of pre-modified noun phrases in simultaneous interpreting from English into Italian: A corpus-based study on EPIC", Serena Ghiselli draws on EPIC to analyze professional interpreters' performances in dealing with a well-known mnemonic challenge when working between non-cognate languages with a reversed lexical order. In "On anaphoric pronouns in simultaneous interpreting" Ana Correia aims to achieve a better understanding of cohesion in interpreter output by looking at anaphoric reference in a language combination not present in EPIC, that is English-Portuguese. Finally, in "Interpreting universals: A study on explicitness in the intermodal corpus EPTIC" Niccolò Morselli sets out to investigate the occurrence of interpreting universals by querying the intermodal corpus EPTIC (European Parliament Translation and Interpreting Corpus, supplementing EPIC with the corresponding written translations) developed at the Department of Interpreting and Translation of the University of Bologna.

Section 2 *Interpreter training: developing and using corpora to train interpreters* opens with the contribution by Andrew Cresswell "Looking up phrasal verbs in small corpora of interpreting: An attempt to draw out aspects of interpreted language". This study relies on data from various existing corpora of English, both as a naturally occurring language and as simultaneous interpreting output to inform the teaching of English phrasal verbs in language lessons in order to improve fluency in non-native trainee interpreters. Next is the contribution by Michela Bertozzi "*Anglintrad*: Towards a purpose-specific interpreting corpus". This study looks at the strategies implemented by simultaneous interpreters and translators working into Spanish when dealing with anglicisms in Italian source speeches. The study adopts an intermodal perspective, that is contrasting interpreted and translated language, with a pedagogic aim. The last paper in this section is "The TIPp project: Developing technological resources based on the exploitation of oral corpora to improve court interpreting" by Mariana Orozco-Jutorán. It shows how leveraging on multilingual corpora, based on genuine court interpreter-mediated interactions, it is possible to build additional resources to improve court interpreting.

This special issue and a parallel edited volume (Russo et al. 2018), both sparked by the First Forlì Workshop on Corpus-based Interpreting Studies held in May 2015 and attended by numerous researchers engaged in this innovative field world-wide, show the geographical spread of the corpus-based approach in interpreting studies. What started as an Italian enterprise centered on the University of Bologna and the University Trieste, has found followers in many European and Asian countries. In this issue, the University of Bologna, the University of Modena and Reggio Emilia (Italy), the University of Minho (Braga, Portugal), the Autonomous University of Barcelona (UAB, Spain) are represented. The parallel edited volume also presents work from other European countries (Poland, Belgium), China and Japan.

We hope these works can serve as an inspiration to other scholars who may join in the effort of creating further language resources such as corpora. These are proving useful in both interpreting and translation research and education as a result of systematic observation afforded by CIS methods.

References

Russo, M., Bendazzoli, C. e Defrancq, B. (eds) (2018) *Making Way in Corpus-based Intepreting Studies*. Singapore: Springer.

"Corpus-based Interpreting Studies: a booming research field", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2304

"Corpora worth creating: A pilot study on telephone interpreting", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2315

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy.

Corpora worth creating: A pilot study on telephone interpreting

By Sara Castagnoli & Natacha Niemants (University of Macerata & University of Modena and Reggio Emilia, Italy)

Abstract & Keywords

English:

This paper reports on the development and use of a corpus of interpreter-mediated phone calls to study features of telephone interpreting (TI) in healthcare settings. After a short introduction on TI and corpus-based studies of remote and on-site community interpreting (CI), the paper discusses ways of exploiting the corpus to analyse interpreters' translation and coordination activities over the phone. It first shows that, notwithstanding some limitations due to data originally collected for non-linguistic purposes, even a small and raw resource can contribute to exploratory analyses of TI, using a qualitative (Conversation Analysis) approach. It then illustrates how opportunities for more systematic research are opened up by corpus annotation. The paper finally reports on some preliminary insights about linguistic and interactional aspects characterizing this type of remote interpreting and makes a tentative comparison with two on-site CI corpora, thereby paving the way to more refined and quantitative investigations.

Keywords: telephone interpreting, community interpreting, interpreter-mediated interaction, healthcare, coordination, remote interpreting

1. Setting the scene: Telephone Interpreting as a sub-type of Community interpreting

Although corpus-based research in the field of interpreting is still lagging compared to Translation Studies, mainly because of the specific challenges involved in the treatment of spoken language data, the last decades have witnessed an increase in the number and variety of interpreting corpora available across modes and settings (see Setton 2011 and Bendazzoli 2015 for a review of such developments, and specific studies in Straniero Sergio and Falbo 2012b). One interpreting modality that has so far escaped empirical, corpus-based research is *remote interpreting* (RI), that is the provision of interpreting services from a distant location using communication technologies (telephone, videoconferencing, web-conferencing); when a telephone line is used to connect the interpreter to some or all of the primary participants, who may be together at one site or at separate locations, RI is usually called *telephone interpreting* (TI), or *over-the-phone interpreting* (Braun 2015).

Demand for RI has been increasing steadily in the last decades, especially in public service settings – such as healthcare and court – that are highly affected by migratory patterns and language issues, and that are normally associated with face-to-face community interpreting (CI). RI's uptake in these contexts can be explained by its several logistical and financial advantages vis-à-vis traditional on-site CI, including a) improved interpreter availability, in terms of language coverage (especially for minority languages) and 24/7 availability with short or no notice, even at peripheral or isolated facilities; b) lower costs, mainly due to savings related to interpreters' travel time; and c) increased confidentiality in delicate situations (see, among others, Ko 2006; Rosenberg 2007; Braun 2015).

The expansion of RI has been accompanied by a number of surveys on user perception and satisfaction, which also brought to the fore some of its perceived disadvantages. According to such surveys, RI is at least as acceptable and effective as on-site interpreting for patients, whereas doctors and interpreters generally show preferences for on-site interpreting over RI, mainly due to the challenges of missing visual information, increased difficulty in building rapport, and possible technical problems (see, among others, Ko 2006; Lee 2007; Locatis et al. 2010; Price et al. 2012). Further studies found that (actual and perceived) difficulties with RI tend to diminish with training and experience, thus pointing to the need for both RI users and providers to be specifically trained to the different challenges of interpreting *in absentia* (see for example Wadensjö 1999; Braun 2006; Kelly 2008; Hlvac 2013).

On the other hand, empirical studies of authentic interactions in this particular communicative situation, aimed to identify typical features of RI as well as factors enhancing or threatening its success, are still scarce. Analysing interpreters' performance in simultaneous RI, for example, Moser-Mercer (2003; 2005) found that interpreting quality deteriorated faster than in on-site performance, possibly because lack of visual presence, among other factors, determined increased stress and an earlier onset of fatigue. Similar findings were reported by Roziner and Shlesinger (2010), who however highlighted considerable discrepancies between objective measures and subjective perception of performance quality. Experimental studies of videoconference-based dialogue interpreting in legal settings conducted in the context of the European AVIDICUS project basically pointed in the same direction, suggesting greater difficulties and a higher cognitive load for interpreters; however, they also found differences in the dynamics of the communication between traditional and video-mediated settings, the latter being characterized by a reduction in the quality of intersubjective relations between participants and greater discourse fragmentation (Braun and Taylor 2012; 2014).

As regards specific research on TI, Wadensjö (1999) compared on-site and telephone interpreting on the basis of two real-life encounters recorded at a police station. She found that the main difference between the two modalities lay in the possibilities they provided for the coordination and synchronization of interaction: telephone interpreting was found to be characterized by different turn length, less overlapping talk and a greater coordination effort on the part of the interpreter, with difficulties also deriving from the lack of visual cues. Based on a larger sample of over

1,000 personally interpreted phone calls, Rosenberg (2007) argued that major difficulties in TI were caused by the lack of a shared frame of reference, but also by the lack of initial briefing, poor sound quality, and unusual turn-taking patterns due to the configuration of the call (speakerphone vs. telephone passing). More promising results are still to come from the ongoing European SHIFT project (Spinolo et al. forthcoming),**[1]** which sets out to provide training in remote CI based on the analysis of authentic telephone and video interpreter-mediated multilingual communication.

2. Corpus-based studies of Community Interpreting

The availability of authentic data on remote interpreting is extremely limited compared with the (slowly) growing number of corpus-based studies of traditional community interpreting, with which RI research shares a number of technical/practical problems and methodological concerns. To quote but a few, a) the difficulty of accessing data and getting permission to use them for scientific purposes, which impacts on corpus design and representativeness, and ultimately on the researchers' objectives (Straniero Sergio and Falbo 2012a); b) the time-consuming nature of data collection and transcription, which limits corpus size and also influences analysis (Niemants 2012); c) the problem of dealing with dialogue-like data including both monolingual and interpreted utterances, where overlaps and other conversational phenomena are hard to annotate or extract automatically (Angermeyer et al. 2012). There is consequently much room for improvement and the few existing CI corpora and corpus-based studies are our closest and most valuable reference.

Research in this field has been mostly qualitative in nature, relying on discourse-analytic and ethnographic methods, and especially on Conversation Analysis (CA; Sacks et al. 1974), which appears particularly well suited for observing interpreting as interaction (see Baraldi and Gavioli 2012; Straniero Sergio and Falbo 2012a; Davitti and Pasquandrea 2014; Dal Fovo and Niemants 2015 for recent overviews). As Meyer and his panelists reminded us at the *Corpus-based Interpreting Studies – The State of the Art* workshop, interaction sequences are usually investigated in detail to pinpoint systematic challenges of community interpreting, for example code switching, dyadic sequences, explanations of technical terms, and to substantiate Wadensjö's (1998) theoretical distinction between "translation" and "coordination" by observing it in professional practice.

As for interpreters' translating activity, a number of independently conducted investigations have shown that turn-by-turn translation is just one of the ways in which interpreters translate the interactions. Research conducted on authentic interpreter-mediated encounters has expanded on the categories of "renditions" identified by Wadensjö (1998) and provided extensive evidence of how interpreters translate and of the reasons why they do it that way (Mason 1999 and 2006; Davidson 2000 and 2002).

The coordinating function of interpreters has also attracted an increasing interest among researchers, leading to the publication of individual and collective endeavours that have further developed Wadensjö's pioneering categorizations. According to Wadensjö, interpreters play their coordinating function when they contextualize their translations in the interaction and when they manage the turn-taking. Coordination is *implicit* when interpreters translate, since the fact of producing a turn in a language implicitly selects the participant who speaks it; coordination is *explicit* when interpreters carry out other actions, which have no counterpart in a preceding original – she calls them "non-renditions" – and which overtly contribute to organizing talk in interaction. Wadensjö herself (1998: 145–151) suggested that the distinction between implicit (through renditions) and explicit coordination (through non-renditions) is not clear-cut, since there exist overlapping areas that can be as interesting as the distinction itself and pave the way to new dichotomies, such as the one introduced by Baraldi and Gavioli:

> We suggest that the distinction posited by Wadensjö between implicit and explicit coordination can be looked at as a distinction between basic and reflexive coordination, both of which can potentially be achieved by renditions and non-renditions. Basic coordination is the smooth achievement of self-reference, without any emergence of problems of understanding and/or acceptance of utterances and meanings. Reflexive coordination is the achievement of self-reference through actions that aim to improve (encourage, expand, implement, etc.), question or claim understanding and/or acceptance of utterances and meanings (2012: 5–6).

The edited volume by Baraldi and Gavioli gives a substantial contribution to research into reflexive coordination activity, as the chapters in it analyse, in different ways and from different perspectives, authentic data from some of the CI corpora available to the scientific community, e.g. the DiK corpus of Portuguese-German and Turkish-German interpreted doctor-patient communication (Bührig and Meyer 2004), the AIM corpus of Italian-English/Arabic/Chinese/French language-mediated interactions in healthcare (Baraldi and Gavioli 2012), and the CorIT corpus of Italian television interpreting (Falbo 2012).

Although CI corpora have so far been analysed qualitatively, their growing size is now opening paths to combine qualitative and quantitative approaches, where corpus technologies can be used to code, count and search existing collections. In this regard, Gavioli et al. (2016) suggested that coding CI corpora should primarily be done to search interesting data for analysis, i.e. to retrieve different types of encounters and interaction sequences, and only afterwards to count what has been searched for, for example lexical items or dyadic vs. triadic sequences. These considerations originate from their use of the AIM corpus, which is probably one of the biggest collections of this kind worldwide.**[2]**

The AIM corpus currently includes about 550 interpreter-mediated medical encounters for over 100 hours audio recording, whose transcripts have been mainly produced using a simple word processor. Although many AIM transcribers still prefer to rely on separate software tools to manage the recordings and produce transcripts, there is a case for using a single interface, where the transcript acts as a dynamic index to the recording. This would allow transcribers to keep symbols to a minimum and facilitate the adding and subtracting of details for the purpose of more precise analyses and of presentations to different types of audiences. Such interfaces include EXMARaLDA, which was used to experimentally audio-link a subset of 19 French-Italian encounters recorded in Italian and Belgian healthcare settings (Niemants 2015), and ELAN, which was tested to audio-link another subset collected for the purposes of a recent project on improved communication in Italian healthcare settings.**[3]** Both EXMARaLDA and ELAN enable researchers, just by clicking on the transcript, to listen to the corresponding audio segment, thereby playing a major role in keeping track of what different transcribers do on the data. Both tools additionally allow for data extraction, enabling one to retrieve lexical matches, like concordances, as well as complex interactional sequences that can be later investigated in greater details.

While being far from the level of annotation added in the Community Interpreting Database,**[4]** the AIM corpus is representative of the efforts that are internationally being made to turn existing CI "spoken corpora", that is 'collections of transcripts of (video)recorded data not included in the corpus' (Straniero Sergio and Falbo 2012a:

31), into "speech corpora", that is multimodal collections where 'the audio and/or video tracks, with the relevant transcripts, are an integral part of corpora themselves' (ibid.: 31–32). As Straniero Sergio and Falbo underline when making such a distinction, 'this difference in the presentation of data decisively affects the analysis potential not only of corpora, but also of the aspects to be investigated' (2012a: 31–32), which is true for both on-site and remote forms of interpreting.

3. Creating and using a Telephone Interpreting corpus

The following sections introduce the work done to create a corpus out of a small existing collection of TI transcripts. Our aim is to show that, considering the paucity of authentic RI data available to the scientific community, even a small, unorthodox corpus can make a valuable – albeit exploratory – contribution to empirical research on this interpreting type.

3.1 Corpus description

Our corpus contains the transcriptions of 30 telephone calls to a remote interpreter made at a healthcare institution of the Emilia Romagna region (Italy) on the occasion of medical encounters involving at least one Italian healthcare provider (doctor, nurse) and a foreign patient with no or limited Italian proficiency. The phone calls were made between 2013 and 2014, when the institution was experimenting with a TI service to assess whether it could effectively be integrated with the existing face-to-face community interpreting service – or replace it in some specific contexts – in order to reduce costs and improve coverage.

The 30 recordings (a sample of about 1/5 of total phone calls made during the testing phase, kept by the external service provider for legal purposes) were listened to and transcribed within the institution to monitor how the TI service was being used, as well as the personnel's attitude towards the service itself. Most transcriptions were provided by some to-be translators during traineeships, mainly for phone calls involving language pairs that they could master (namely, Italian plus English/French/Albanian/Polish); some recordings involving Chinese or Arabic interpreters were also transcribed leaving out non-Italian turns, based on the assumption that even partial transcriptions would be sufficient for the non-linguistic quality control envisaged. For the same reason, transcribers were not provided with specific guidelines for the transcription nor for the annotation of paralinguistic features such as pauses, hesitations or overlaps. Plain, orthographic transcriptions were produced using a word processor. Access to the recordings was subject to a cumbersome procedure which involved asking permissions to listen to individual audio files; these would remain accessible for 4–5 days and could not be downloaded.

Lack of access to the original recordings implies that, when turning these texts into a corpus, we had to accept some major limitations as "irreparable". The quality of transcripts stands out as the most problematic aspect: a) some of them are incomplete, as they involved turns in languages unknown to transcribers; b) as they were produced by different translators, not specifically trained for the purpose and without any guidelines, inconsistencies – and even inaccuracies – are inevitable both within and across transcripts, especially as regards the annotation (when provided) of paralinguistic features. In addition, the lack of relevant audio tracks prevents any development of this *spoken corpus* into a multimodal *speech corpus* (see section 2).

The nature of the data directs and constrains the range of possible research objectives. On the one hand, the above-mentioned limitations affect the analysis potential of the corpus by preventing research on aspects that are specific to spoken language data, like phonetics, prosody, non-verbal features and so on. On the other hand, the small size of the corpus – about 22,500 total words – rules out the possibility to carry out statistical analyses of lexical frequencies. Moreover, the sample comes from a single source and is not sufficiently large to generalise. We believe, however, that by highlighting the recurrence of given lexical and interactional phenomena, the corpus can still be a precious resource to start identifying typical features of interpreting in this particular communicative situation.

3.2 Some insights obtained through qualitative investigations of the raw collection

Our first approach to the corpus was through manual, qualitative investigations of the transcripts, using conversation analysis (CA) methods. Despite the small size of the corpus, it was possible to identify common courses of interaction and recurring features of participants' verbal behaviour, mainly connected to the lack of visual information and shared contextual knowledge, and to observe their positive or negative impact on the success of telephone-interpreted phone calls. The main findings of this study (fully reported in Niemants and Castagnoli 2015) are summarised in the following paragraphs.

The analysis of transcripts shows that conversations in the corpus seldom follow the turn-taking sequence Speaker 1 – Interpreter – Speaker 2 – Interpreter – Speaker 1 – Interpreter and so on: while some examples of this pattern can indeed be observed within individual phone calls, the latter are essentially structured as sequences of monolingual dyadic exchanges between the interpreter and one of the primary participants (see Jefferson 1972 on *side sequences* in general, and Kelly 2007 on *side conversations* or *side talk* in the context of TI), which may or may not be subsequently summarized by the interpreter into the other language for the benefit of the excluded party (see Merlini 2015). Although, according to existing standards of practice, side conversations should be avoided by telephone interpreters (Kelly 2007: 118), the analysis of transcripts suggests that they play a fundamental role in establishing shared ground on which the conversation can then continue. In particular, corpus data suggest that an initial dyadic briefing between the healthcare provider and the interpreter – during which the former provides some basic information about the patient and the reasons for the encounter, and also informs the interpreter about how the conversation is to take place (for example if there is a speakerphone) – is essential to provide contextual information which would be taken for granted in on-site encounters but which is missing to remote interpreters, and may reduce the need for extensive negotiation afterwards.

Corpus data indicate that the lack of a 'shared frame of reference' (Rosenberg 2007: 75) between the primary participants – sitting together at the same location – and the remote interpreter determines knowledge asymmetries at several levels. On a macro-level, remote interpreters are not only physically absent, but they may also not be familiar with the primary participants' local reality (towns, hospital/medical facilities, proper names and so on). Even more significantly, remote interpreters lack information about things that are happening in the room where the medical encounter takes place, including on-going non-verbal communication. The lack of visual clues, which several authors have described as the most problematic and stressing feature of TI (see section 1), has a negative impact on turn management and represents a major limitation in settings where practical information needs to be conveyed.

As far as turn management is concerned, the corpus contains several occurrences of interruptions and uncertainty about who is entitled to the next turn. While in face-to-face situations non-verbal behaviour such as gestures, posture, mimics and gaze have a role in guiding the interaction (see, among others, Wadensjö 1999: 254), in TI interpreters do not have access to these clues and need to rely on explicit verbalisations by the interlocutors (utterances like *Adesso te la passo così glielo dici* 'Now I'll put you through so you tell her', *Aspetta che te lo passo* 'Wait I'll put you through', *Adesso te la ripasso* 'Now I'll put you through again' are indeed common in the corpus). Difficulties in turn management also determine the presence of long, uninterrupted turns especially on the part of healthcare providers, who tend to accumulate (even unnecessary) information or questions before giving the floor to the interpreter. The major risk of such information overload is for interpreters to miss important details, which entails requests for clarifications and repetitions, and more extensive negotiation in general. Corpus data thus seem to confirm Wadensjö's remark that translating and coordinating the talk exchange is more complicated for interpreters in TI than in face-to-face interaction, so participants to a telephone-interpreted encounter should 'make a special effort to express themselves clearly and verbalize any non-verbal activities that may have an impact on the ongoing interaction' (1999: 262).

Overall, the corpus provides evidence that the success of TI may depend on the specific healthcare setting involved and the type of information to be conveyed. As suggested in previous literature (Villarruel et al. 1999: 268; Price et al. 2012), TI is generally smooth and effective in settings in which routine information is exchanged. It is the case, for instance, of interviews preceding paediatric vaccinations, whose contents are highly predictable, as questions and answers normally focus on the child's health, on reactions to previous vaccinations and on informed consent. On the contrary, TI turns out to be less effective in "educational" scenarios in which practical information needs to be conveyed and visual clues are virtually indispensable for mutual comprehension (as also observed by Price et al. 2012), as in the case of one interpreter being asked to translate instructions on how to perform an insulin injection; or when the patient is not cooperative, and building rapport through TI becomes more complicated.

In sum, qualitative analyses proved useful to identify recurrent actions and verbal behaviours which can have an impact on the success of telephone-mediated encounters, and ultimately point to the need to develop the participants' awareness of the additional challenges of TI compared to on-site interpreting.

3.3 Annotating the data for more refined, corpus research

In an attempt to enable more systematic investigations as well as comparative analyses (see section 4), we tried to turn the collection of transcripts into a "proper" corpus which can be searched with corpus-based techniques.

Transcripts were anonymised (by substituting personal and geographical names) and standardised as much as possible in their format. Original text files were converted into xml documents; each phone call is enclosed within a <text> element, with a unique identifier, and speakers' turns (possibly containing several utterances) are encoded as separate <s> elements, progressively numbered.**[5]** This allowed us to add annotations at different levels, in order to encode (implicitly) available extralinguistic information and make it searchable: while we cannot aim at any quantification of interactional aspects, given the size of the corpus and the limited control we had over transcripts, our goal is to be able to make more focused searches as well as retrieve all instances of given elements to be then analysed qualitatively. Because of the impossibility to access original recordings, we decided to annotate only information and features that required the least interpretive effort (see Angermeyer et al. 2012).

Basic categories include descriptive metadata about the date and the setting in which phone calls took place (mainly local vaccination centre or hospital department, such as Maternity, Obstetrics and gynaecology unit, A&E and so on). This information was derived from official reports and coded in a sort of "header" at the <text> level, together with details about the language pair involved and the transcriber's name. Additional descriptive metadata is provided at the turn level, where we annotated speaker's role (doctor/patient/interpreter) as well as the language(s) involved. Basic linguistic annotation, namely Part-of-Speech tagging and lemmatisation, was also introduced; however, as the corpus could only be processed as a single, monolithic entity – it is a single file, comprising parts written in different languages – and since POS taggers can normally handle one language at a time, we decided to treat it as if it were a monolingual Italian corpus (Italian being shared by all interactions within the corpus). Consequently, turns in other languages are not properly tagged.

This basic level of annotation can be leveraged to investigate a number of interesting aspects, for example from differences in interaction patterns across healthcare settings to interpreters' "formulations" of previous turns at talk (Baraldi and Gavioli 2010; Baraldi 2012). For instance, example (1) shows selected concordances obtained by searching for the lemma *dire* "to say/tell" in interpreters' turns, and suggests that the verb is often used within turns directed to healthcare providers containing either their renditions of the patient's utterances (as in cases *a* and *b*) or summaries of what they have just told the patient in a different language (*c* and *d*).**[6]**

(1)

(a) 676: <s_speaker int><s_trans rendition>: Eh , infatti **<dice>** che non è ... non è un infortunio .

Uh, he says that it's not... it's not an accident .

(b) 3866: <s_speaker int><s_trans rendition>: Ha <**detto**> che ogni tanto la bimba soffre di mal di pancia . Dolori lievi .

He said that sometimes the girl has got stomach ache. Mild pains .

(c) 1166: <s_speaker int><s_trans non-rendition>: # sì , gli ho **<detto>** della farmacia che si deve informare # al pronto soccorso .

 # yes , I told him about the pharmacy that he needs to get information # at the emergency room .

(d) 1395: <s_speaker int><s_trans non-rendition>: Sì . Sì . Le ho **<detto>** che deve portare il referto al pediatra per i controlli successivi .

Yes . Yes . I told her that she has to bring the medical report to the pediatrician for the next checkups.

We decided to additionally include two types of analytical annotations which – according to Angermeyer et al. (2012) – are usable with community interpreting data in general (thus enabling comparisons as those provided in

section 4) and do not require much subjective interpretation (thus limiting decision-making at the annotation stage). To start with, we annotated the language of turns distinguishing monolingual (*unmixed*) and multilingual (*mixed*) turns: in community settings it is quite frequent for participants and interpreters to produce mixed utterances, as a result of code-mixing, code-switching or ad hoc borrowing (ibid.: 288–289; see also Anderson 2012; Meyer 2012), and it is arguably worth investigating whether the same occurs in TI. Turns were annotated as *mixed* whenever some kind of language mix – from a single lexical item to a longer code-switch – took place, so that a search for *mixed* turns would retrieve occurrences of any type of bilingual speech. The 75 *mixed* turns in our corpus (out of 1605 total turns) are interpreters' turns (with one exception, discussed in section 4.1); the large majority of such turns correspond to a change in the primary interlocutor, as exemplified in (2),**[7]** with only a few cases of real code-mixing (as in 3).

(2)

<s n="40" speaker="doc" langstat="unmixed" trans="original" lang="it"> Sì, va bene. Senta, potrebbe chiederle se l'allatta ancora al seno e se dalla nascita ad oggi ha avuto nessuna malattia? Sul bambino. Gliela passo. </s>

Yes, fine. Listen, could you ask her whether she's still breastfeeding him and whether he's had any illnesses since birth? About the baby boy. I'll put you through.

<s n="41" speaker="int" langstat="mixed" trans="rendition" lang="it-al"> **Va bene. Grazie. [***((in Albanian))* **Allora, signora. Come terza domanda il dottore vuole sapere se lei ha allattato il bambino al seno dalla nascita e continua tuttora?]** </s>

Ok. Thanks. [((in Albanian)) So, madam. As a third question the doctor wants to know whether you have been breastfeeding the baby since birth and whether you still continue to?]

(3)

<s n="55" speaker="int" langstat="mixed" trans="rendition" lang="it-en"> Hello. **The lady ha detto..** said that when you feel ill you have to eat something sweety and everything that you feel ill you have to eat something sweet to... okay? </s>

The second level of analytical annotation added to the corpus, which is called "translation status", is more interpretive in nature (it requires that the researcher actually looks at the data to make decisions) and is based on Wadensjö's (1998) classification of interpreters' utterances as *renditions* vs. *non-renditions*. This annotation is meant to record whether interpreters' turns correspond to translations of prior utterances made by the primary participants (whose turns are, consequently, always annotated as *original*), or do not have any counterpart in a preceding original turn, thus pointing to instantiations of interpreters' coordinating role. Identifying source-target pairs in dialogue interpreting is not a trivial task, not only because the extent to which interpreters' renditions relate to original utterances may vary (see Wadensjö's more refined categories of renditions), but also because source-target turns may not be adjacent. For example, in (4), the interpreter's turn n= "24" is arguably a rendition of two previous turns by the doctor (namely n= "11" and n= "17"), from which it is separated by a number of non-renditions, consisting mainly in (requests for) clarifications.

(4)

<s n="11" speaker="doc" langstat="unmixed" trans="original" lang="it"> Allora, dovrei dire alla mamma che **deve fare due iniezioni. Una sulla coscia destra, che contiene un attivo contro difterite, tetano, pertosse, epatite B, poliomelite ed emofilo B.** </s>

So, I should tell the mother that he has to get two injections. One on the right thigh, which contains an active ingredient against diphtheria, tetanus, whooping cough, hepatitis B, polio and haemophilus influenzae B.

<s n="12" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Mi scusi, una sulla coscia ? </s>

I'm sorry, one on which thigh?

<s n="13" speaker="doc" langstat="unmixed" trans="original" lang="it"> Destra. </s>

The right one.

<s n="14" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Okay, contro la difterite. </s>

Okay, against diphtheria.

<s n="15" speaker="doc" langstat="unmixed" trans="original" lang="it"> Non solo la difterite. Anche tetano, pertosse, epatite B, poliomelite ed emofilo B. </s>

Not only diphtheria. Also tetanus, whooping cough, hepatitis B, polio and haemophilus influenzae B.

<s n="16" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Okay, okay. </s>

<s n="17" speaker="doc" langstat="unmixed" trans="original" lang="it"> **Mentre nella coscia sinistra farà una puntura che contiene il vaccino antipneumococcico**. </s>

Whereas on the left thigh he will get an injection which contains the vaccine against pneumococcus.

<s n="18" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Anti? </s>

Against?

<s n="19" speaker="doc" langstat="unmixed" trans="original" lang="it"> Antipenumococco. </s>

Against pneumococcus.

<s n="20" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Okay. </s>

<s n="21" speaker="doc" langstat="unmixed" trans="original" lang="it"> Le passo la madre. </s>

I'll put you through to the mother.

<s n="22" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Okay, va bene. Grazie. </s>

Okay, fine. Thank you.

<s n="23" speaker="paz" langstat="unmixed" trans="original" lang="fr">Allo? </s>

Hello?

<s n="24" speaker="int" langstat="unmixed" trans="rendition" lang="fr"> Allo, bonjour. Alors, **le médecin dit qu'il va faire deux injections au bébé, une dans la cuisse gauche, et c'est contre la blessure (XXX) et l'hépatite. Et il va lui faire une autre injection dans la cuisse droite, et c'est contre la méningite.** Halo, vous m'entendez ? </s>

Hello, good morning. So, the doctor says that he will give two injections to the baby, one on the left thigh, and it's against the injury (XXX) and hepatitis. And he will give him another injection on the right thigh, and it's against meningitis. Hello, can you hear me ?

Considering that we can only establish correspondences at the level of speakers' turns (rather than more fine-grained utterances), we decided to annotate as renditions only turns containing a significant amount of propositional elements that are retraceable to something which was previously said by other participants. In other words, interpreters' turns including both translations of prior utterances and other coordinating activities were annotated as non-renditions whenever the latter were more numerous than the former.

Even so, the translation status of some turns is not easily determined. This is the case of turns like examples (1c) and (1d) above, where the interpreter summarizes for the Italian doctor what has just been negotiated with the foreign patient. Although these kinds of turns admittedly render talk for a speaker who was temporarily excluded, the corresponding *original* is the interpreter's turn and not a stretch of talk by a primary participant; as a consequence, strictly following Wadensjö's classification, these turns were treated as *non-renditions*.

The situation is possibly even less clear-cut when dealing with interpreters' replies to patients' requests for clarification, as in (5) below. A diabetic patient is receiving instructions on how to measure her blood sugar. Following the interpreter's rendition of the doctor's utterance, the patient asks for two clarifications on the content just rendered, and the interpreter responds with two turns in which he replies autonomously, albeit recalling some content he knows – and has already rendered – because the doctor first voiced it.

(5)

<s n="58" speaker="paz" langstat="unmixed" trans="original" lang="en"> In a week I control three times in a day? </s>

<s n="59" speaker="int" langstat="unmixed" trans="non-rendition" lang="en"> **No, no. All morning you have to control**. </s>

<s n="60" speaker="paz" langstat="unmixed" trans="original" lang="en"> All morning in every day. In the morning. </s>

<s n="61" speaker="int" langstat="mixed" trans="non-rendition" lang="it-en"> **Sì. You have to control in the morning.** </s>

From a certain viewpoint, the above interpreter's turns could be annotated as renditions – especially if more fine-grained categories were used, e.g. Wadensjö's "multi-part renditions" for multiple interpreting utterances corresponding to one original. We decided, however, to treat them as non-renditions, in order to show that rendering originals is not enough to ensure participants' mutual understanding, and that interpreters thus need to do more, such as recalling and repeating something that has already been said and translated, to encourage and promote understanding and participation. A more detailed analysis of the possible nature of non-renditions is provided in section 4.2.

4. Telephone vs. on-site interpreting – Comparing corpora and discussing results

Notwithstanding the corpus' small dimensions and the problems raised by some of the annotations added, the analyses in section 3 have already shown some of their potential. We believe, however, that the comparison with existing on-site CI corpora can yield even more interesting insights. For the purposes of this paper, we will use as reference corpora the two time-aligned subsets of the AIM corpus that we have personally transcribed, namely a small collection of interpreter-mediated interactions in Italian and Belgian healthcare settings (Niemants 2015) and a recently collected corpus of mediated and non-mediated doctor-patient interactions in the Emilia Romagna region (see footnote 3). These were transcribed using two different multi-tier transcription tools – EXMARaLDA and ELAN – and will henceforth be referred to as "the EXMARaLDA sub-corpus" and "the ELAN sub-corpus", respectively.

4.1 A closer look at mixed turns

As we have seen above, mixed turns in our TI corpus amount to 75. Either by refining the quantitative search for speakers or by having a qualitative look at single occurrences, it appears that all the matches except one are in the interpreter's turns. In other words, with the exception of example (6) below, where the Italian doctor utters an *all right?* in English, doctors never switch code in our TI corpus.

(6)

<s n="1" speaker="doc" langstat="mixed" trans="non-rendition" lang="it-en"> Sì. Ha capito tutto. **All right?** Okay. Adesso mi deve dire, che non succede, però se dovesse, se ha delle crisi ipoglicemiche, che significa un

abbassamento della glicemia, lei inizia a sudare, ad avere tremori, crampi, un malessere generale, deve prendere subito dello zucchero. Okay? E poi mangiare. Okay? </s>

Yes. She understood everything. All right? Okay. Now you have to tell, this doesn't happen, but in case it happened, if she has hypoglycaemic reactions, which means that her blood glucose level falls, she starts sweating, having tremors, cramps, general sickness, she must have some sugar immediately. Okay? And then eat. Okay?

The absence of mixed turns in doctors' talk might not be surprising if our TI corpus was only analysed on its own: it may simply indicate that doctors fully delegate the translation activity to interpreters – this is what they are contacted for in the first place – and never switch code to communicate with the patients who are sharing their physical space. While we cannot speculate on what precedes and follows the phone calls, we can state that in our corpus doctors never try to address patients directly during the call, and wonder whether this practice has some implications for the role of the telephone interpreter. As Mason (2006), Zorzi (2012) and Gavioli (2015), among others, have been showing for on-site CI, the interpreter's role is not fixed, but highly depends on the actions of other participants in the interaction, which may arguably be the case for telephone interpreting, too.

The absence of code-switches in doctors' turns, however, becomes more revealing if we compare our TI corpus with reference on-site CI corpora, which both contain instances of healthcare workers addressing patients directly.

Example (7) is taken from the "EXMARaLDA sub-corpus", where code-switches were explicitly annotated as such and can easily be retrieved using EXMARaLDA concordance tool EXAKT. Here a female Italian doctor (Doc) utters two words in French, asking the patient to breath-in through his mouth ('open mouth'). This piece of information is acknowledged by the interpreter (Int), who utters the acknowledgment token *ah* in partial overlap with the doctor's turn. She then redesigns this turn to make sure the patient follows the doctor's instructions ('ah okay good you breathe through your mouth').

(7)

Doc: bouche [ouverte]

mouth [open]

Int: [ah] okay bon tu respires avec la bouche

[ah] okay good you breathe through your mouth

Example (8) is taken from the "ELAN sub-corpus", where language status was explicitly codified as "Italian" (when turns are entirely uttered in this language), "non-Italian" (when turns are uttered in foreign languages, here mainly Arabic, English and French), "mixed" (when turns contain propositional contents in both Italian and one or more foreign language(s)) and "international" (when speakers utter minimal responses that are hardly falling within the other three categories – such as *mm hm*, *okay* and the like – or when they produce non-verbal vocalization like *laughter*, *cough* and the like), and where healthcare providers (mainly mid-wives) often switch code to address patients directly. Here the mid-wife partially self-translates what she has just said in Italian, for the patient (Paz) to understand what is about to happen (pressure measurement). In other cases, mid-wives retrieve foreign words uttered by the patients and integrate them into their Italian turns at talk (e.g. *te la do adesso la drugs sì*, 'I'll give it to you now the drugs yes').

(8)

Doc: ti provo la pressione [la tension]

I'll measure your pressure [your pressure]

Pat: [oui d'accord]

[yes, fine]

The tendency of healthcare workers to switch code in order to address patients directly is confirmed by other researchers working on bigger corpora of on-site CI in healthcare, such as Meyer (2012) analysing data from two projects on *ad hoc* interpreting for Turkish and Portuguese patients in hospitals in Hamburg, and Anderson (2012) studying a subset of the AIM corpus including out-patient visits with English speaking patients. Both authors show that primary participants can have some level of proficiency in their respective languages and try to communicate directly, which poses problems of coordination for interpreters and inevitably affects their role. More precisely, if primary speakers are able to understand and talk to each other, interpreters may be called to stop translating, to stay on a 'stand-by-mode', as Angermeyer suggestively describes it (2008: 391), and to monitor participants' understanding in order to decide when it is time to move in – because they do not understand each other – and out – because they manage to communicate directly – of the conversation.

The near absence of doctors' mixed turns in our small TI corpus might be due to the fact that the patients in the corpus mainly speak languages that are unknown to doctors, but might also suggest that telephone interpreters are not called to 'monitor and mould' code-switching and other 'participant behaviours that can potentially index an (at least partial) understanding of the 'other' language on the part of one or more primary participants' (Anderson 2012: 144). This hypothesis obviously needs to be tested using bigger TI and CI corpora, where translating and coordinating activities such as monitoring can be qualitatively and quantitatively compared. However, our tentative comparison raises an interesting question about the potentially different nature of on-site and remote interpreting, namely: does the absence of direct communication between doctors and patients project a mainly translating role for telephone interpreters? In other words, does their interpreting activity on the phone mainly consist in rendering primary speakers' talk?

4.2 A closer look at non-renditions

The number of renditions and non-renditions in our TI corpus (namely 207 vs. 584) in fact suggests that telephone interpreting does *not* mainly consist in rendering primary speakers' talk but involves a high dose of coordinating activities. This preliminary result is in line with recent studies on dialogue interpreting (see Baraldi and Gavioli

2012; Dal Fovo and Niemants 2015 for two recent collections) where, irrespective of the languages spoken, interpreters appear to do much more than translating.

Starting from the assumption that both on-site and remote interpreting consist in translating (mainly through renditions) *and* coordinating (mainly through non-renditions) work, we will now compare some non-renditions retrieved from our TI corpus with some of those recurring in our two reference corpora. Our objective is to see whether there are any qualitative differences between non-renditions in on-site and telephone interactions, and to make hypotheses on their possible implications for TI practice and training.

If we have a closer look at turns annotated as non-renditions in our TI corpus, it appears that many instances have to do with the medium and the interpreting service provided through it. This is the case of utterances such as *pronto?* (the Italian "hallo" when answering the phone) or of utterances having the following structure: 'good evening (or good morning), interpreter for the Chinese (or other) language' (as in example (9)).

(9)

<s n="1" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Buonasera, interprete di lingua cinese. </s>

Good evening, interpreter for the Chinese language.

A number of non-renditions also have to do with meaning negotiation: they are used by interpreters to (dis)confirm that they are hearing and/or understanding properly, thereby signalling that the other speaker can(not) go on speaking. This is often done through minimal responses such as *sì*, *okay*, *mm hm* (see turns n= "8", "10", "11" and "13" in example (10)).

(10)

<s n="8" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Va bene, okay. E poi?</s>

Fine, okay. And then?

<s n="9" speaker="doc" langstat="unmixed" trans="original" lang="it"> E poi se ha # allergie # alle medicine o agli alimenti.</s>

Then if he has # allergies # to drugs or food.

<s n="10" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> # poi # sì</s>

 # then # yes

<s n="11" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Okay, sì, ha #</s>

Okay, yes, he has #

<s n="12" speaker="doc" langstat="unmixed" trans="original" lang="it"> # se ha avuto reazioni con i vaccini fatti finora.</s>

 # if he had reactions with the vaccines he's been given so far.

<s n="13" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Okay, reazioni... va bene, okay. </s>

Okay, reactions… fine, okay.

<s n="14" speaker="doc" langstat="unmixed" trans="original" lang="it"> Okay.</s>

<s n="15" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Sì? Ehm... posso parlare già con la mamma?</s>

Yes? Ehm… can I already speak to the mum?

<s n="16" speaker="doc" langstat="unmixed" trans="original" lang="it"> Sì, sì sì # parli</s>

Yes, yes yes # speak

As the example above shows, and Gavioli (2012) has also pointed out, *okay* systematically occurs when the translation is about to take place and plays the double role of (1) showing understanding of what has just been uttered by one primary speaker (for example the first *okay* in turn n= "13", *Okay, reazioni… va bene, okay*) and (2) projecting the beginning of the translation for the other (for example the second *okay* in that same utterance). But the doctor does not understand that the translation is about to start and utters himself an *okay*, which requires a greater conversational effort on the part of the interpreter, who explicitly asks whether she can start translating for the mother and waits for the doctor to answer before doing so (in turns omitted here).

In addition to meaning negotiation, many non-renditions serve to summarize the gist of preceding turns and correspond to the formulations we problematized in section 3.3, which often contain the lemma *dire* followed by what has actually been said by primary participants and/or interpreters.

(11)

<s n="62" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Dottoressa, eccomi. Lei ha detto che... le ho spiegato che c'è un costo e che non è necessario che lei debba stare per una semplice medicazione per giorni e giorni in ospedale. Possono fare un'eccezione due giorni ma non di più. Stop. Lei ha detto che stasera parlerà con il marito e poi provvederanno... le ho anche suggerito di andare nella parrocchia di appartenenza, di zona di residenza se c'è qualche volontario che, lo fanno di solito, la può accompagnare. Tutto qui. Io ho finito il mio lavoro. </s>

Doctor, here I am. She said that… I explained to her that there's a cost and that it's not necessary that she stays for a simple dressing for days and days in hospital. They can make an exception for two days but no longer than that. That's it. She said that tonight she's going to talk to her husband and they will take action… I also suggested her to go to the parish they belong to, where they live if there's some volunteer that, they usually do it, can accompany her. That's it. I've finished my job.

The interpreter here summarizes the gist of the previous turns in two ways: she first recalls what she has just told the mother (*le ho spiegato che*, "I explained to her that") and then formulates what the mother has actually said (*lei ha detto che*, "she said that"). Both summaries are addressed to the doctor, who is also informed of what the interpreter has autonomously suggested, that is to go to the local parish and ask for a volunteer to accompany her, as they usually do.

Some non-renditions finally have to do with the interpreter's need to take time and write things down. Utterances like *un attimo che scrivo* ("wait a moment I'm writing") make it explicit that there is a difficulty in dealing with long and dense doctors' turns, and arguably point to the importance of note-taking in telephone interpreting.

(12)

<s n="1" speaker="int" langstat="unmixed" trans="non-rendition" lang="it"> Un attimo che scrivo. Per 21 giorni, una al giorno.</s>

Wait a moment I'm writing. For 21 days, one a day.

If we now search for non-renditions in our on-site reference corpora, we are confronted with similar and divergent courses of action. Starting from similarities, both the EXMARaLDA and the ELAN sub-corpora also contain meaning negotiation sequences where minimal responses play a major role in showing understanding of what has been previously said or negotiated and signalling the transition to translation. This is the case in (13), where the interpreter (Int) first acknowledges receipt of what the Italian speaking patient (Paz) has just said (*va bene (.) okay*), then uses the same token *okay* – but with a French pronunciation – to address the healthcare worker.

(13)

Pat: no

Int: va bene (.) okay

fine (.) okay

(2)

Int: okay donc

okay then

In the ELAN corpus, *okay* is very widely used as a feedback token: when uttered with a rising intonation, be it on its own or at the end of longer turns at talk, it generally checks for the patients' understanding and is often followed by their confirmation; when uttered with a falling intonation, usually on its own or at the beginning of longer turns, it shows understanding and signals the transition to translation. The selected example is representative of the latter case, where the interpreter utters three *okay* in less than four seconds to acknowledge receipt of what the Arabic patient has just said and start rendering it into Italian for the midwife, who had explicitly invited the interpreter to tell her.

(14)

Int: [okay però]

[okay but]

Pat: [hata] khti [lama] bitihmil

[even] my sister [when] she got pregnant

Int: [okay] okay dice che loro in famiglia…

[okay] okay she says that in her family…

Both the EXMARaLDA and the ELAN sub-corpora also contain formulations that summarize the gist of previous turns at talk and that are similarly pre-faced by expressions like *ti dice che…* or *ti ha detto che…* (meaning "s/he tells" – or "told" – "you that…"), but they additionally present a number of non-renditions we have not found in our TI corpus, such as directions on how to reach local healthcare facilities or clarifications and explanations on routine examinations the patient shall undergo.

Example (15) is taken from the EXMARaLDA sub-corpus, where direction-giving sequences were explicitly annotated as "instruction" among departures from the traditional triadic sequence organization in interpreter-mediated interactions. Following the midwife's indication to go to a certain office, the interpreter provides the patient with all the necessary directions to reach that place alone, going so far as to write them down on a sheet of paper she can later use as an *aide-mémoire*.

(15)

Int: [scri-] écris si c'est écrire ça ehm allora (2) je te l'écris ici (.) via Mandorla (1) ehm autobus (1) numéro deux (.) pour aller (2) ça c'est l'autobus qui porte à Modena

[writ-] write if it's writing this ehm so (2) I'm writing this here for you (.) via Mandorla (1) ehm bus (1) number two (.) to go (2) this is the bus that goes to Modena

Example (16) comes from the ELAN sub-corpus and can be retrieved using the software multiple layer search, which allows one to investigate interactional patterns such as Patient-Interpreter-Midwife and to explore the translation status of interpreters' turns. On a closer look, while some instances unsurprisingly are Italian renditions of what foreign patients say, followed by the midwife reception, many interpreters' turns are uttered in the patients' language and are thus non-renditions playing a wide range of functions, for example providing feedback, giving directions, expanding explanations, and asking for or making clarifications, as is the case here, where the interpreter clarifies that the pap test would not have taken longer if it had been performed.

(16)

Int: w chufi (.) kun 'amlatu lik (.) kun 'amlatu nafs el waqt gha diri gha tghulik safi ['amaltu] (.) lahaqach had li bghat dakhlatu kant gha tamsah- ma'mlatuch hit 'ank des pertes ktira

and look that (.) if she had done that (.) it would have taken the same amount of time immediately she would have [told] you I'm done (.) because in the end she had already inserted- she didn't do it because you have heavy discharges

Clarification, explanation, and direction-giving are activities that the interpreter is more or less explicitly delegated to carry out on behalf of the healthcare staff, which again is not an isolated phenomenon, as the tendency to delegation is confirmed by analysts working with other corpora (such as different subsets of the AIM corpus: see Baraldi 2009; Gavioli 2015) and other methods (such as participant observation and interviews: see Hsieh 2010).

Given the near absence of these activities (both as delegated by healthcare workers and self-initiated by interpreters) in our small TI corpus, we can make the hypothesis that telephone interpreters are not called to play a role that has been variably and arguably labelled as co-interviewer (Davidson 2000), co-diagnostician (Hsieh 2007), co-therapist (Bot and Verrept 2013). Again, this preliminary result should be verified in bigger corpora, where quantitative approaches cannot do without the qualitative explorations that enable one to go deep into the nature of interpreter-mediated interactions and of interpreters' contributions to them. But as limited as it may be, our tentative comparison has the merit of exploring two possible research directions and of showing their implications for telephone interpreting users and providers.

5. Conclusions

This study set out to provide data-based reflections on telephone interpreting, in order to start filling the gap in empirical research about this particular interpreting type, which is increasingly common in some community settings. Starting from the widely-shared assumption that community interpreters both translate and coordinate the interaction, the results of our research suggest that TI is characterised by some linguistic and interactional specificities which distinguish it from on-site CI, and which are largely determined by interpreters' physical and experiential remoteness as well as by the lack of visual information that the medium entails.

Corpus data indicates that in TI primary speakers do not try to communicate directly: this may suggest that the monitoring role found in on-site CI could be irrelevant in its remote forms, where interpreters are called for translating and expected to do primarily this. These preliminary findings thus have significant implications for interpreters' training, as would-be-interpreters should be aware of the different roles they may be expected to play, and of how the actions of primary participants can affect their activity.

The study also highlights the need to raise healthcare providers' awareness of the peculiarities of TI, where the lack of a shared frame of reference requires adaptation of habitual on-site CI practices. For instance, some delegations that healthcare providers often make during on-site encounters are problematic in TI, mainly because interpreters may not be familiar with the local reality, and cannot therefore fulfil the same facilitating function that they are usually charged with in on-site CI. Healthcare providers should also be alerted to the fact that the lack of visual information entails a greater negotiation effort, thus requiring more thoughtful communicative behaviour.

From the point of view of interpreting research, the study confirms the worthiness of transcribing, annotating and analysing even the smallest collections of authentic (telephone/remote) interpreting data, as these can provide invaluable exploratory insights, encouraging and justifying the creation of more full-fledged corpora. In particular, further research in the field of RI would evidently benefit from the availability of multimodal corpora where transcripts are linked to original audio or video recordings, as the joint analysis of the two types of data can provide better descriptions of complex speech patterns and phenomena than is possible on the basis of transcripts alone.

References

Anderson, Laurie (2012) "Code-switching and Coordination in Interpreter-mediated Interaction" in *Coordinating Participation in Dialogue Interpreting,* Claudio Baraldi and Laura Gavioli (eds), Amsterdam/Philadelphia, John Benjamins: 115–48.

Angermeyer, Philipp S. (2008) "Creating Monolingualism in the Multilingual Courtroom", *Sociolinguistic Studies* 2, no. 3: 385–403.

Angermeyer, Philipp S., Bernd Meyer, and Thomas Schmidt (2012) "Sharing Community Interpreting Corpora" in *Multilingual Corpora and Multilingual Corpus Analysis,* Thomas Schmidt and Kai Wörner (eds), Amsterdam/Philadelphia, John Benjamins: 275–94.

Baraldi, Claudio (2009) "Forms of Mediation: the Case of Interpreter-Mediated Interactions in Medical Systems", *Language and Intercultural Communication* 9, no. 2: 120–37.

Baraldi, Claudio (2012) "Interpreting as Dialogic Mediation: The Relevance of Expansions" in *Coordinating Participation in Dialogue Interpreting,* Claudio Baraldi and Laura Gavioli (eds), Amsterdam/Philadelphia, John Benjamins: 297–326.

Baraldi, Claudio, and Laura Gavioli (2010) "Interpreter-mediated Interaction as a Way to Promote Multilingualism" in *Multilingualism at Work: From Policies to Practices in Public, Medical and Business Settings*, Bernd Meyer and Birgit Apfelbaum (eds), Amsterdam/Philadelphia, John Benjamins: 141–62.

Baraldi, Claudio, and Laura Gavioli (eds) (2012) *Coordinating Participation in Dialogue Interpreting*. Amsterdam/Philadelphia, John Benjamins.

Bendazzoli, Claudio (2015) "Corpus-based research" in *Routledge Encyclopedia of Interpreting Studies*, Franz Pöchhacker (ed.), London/New York, Routledge: 87–91.

Bot, Hanneke, and Hans Verrept (2013) "Role Issues in the Low Countries, Interpreting in Mental Healthcare in the Netherlands and Belgium" in *The Critical Link 6. Interpreting in a changing landscape,* Cristina Schäffner, Krzysztof Kredens, and Yvonne Fowler (eds), Amsterdam/Philadelphia, John Benjamins: 117–31.

Braun, Sabine (2006) "Multimedia Communication Technologies and their Impact on Interpreting" in Mary Carroll, Heidrun Gerzymisch-Arbogast, and Sandra Nauert (eds), *Audiovisual Translation Scenarios*. Proceedings of the Marie Curie Euroconference "MuTra: Audiovisual Translation Scenarios", Copenhagen, 1-5 May 2006, URL: **http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Braun_Sabine.pdf** (accessed 11 November 2016).

Braun, Sabine (2015) "Remote Interpreting" in *Routledge Handbook of Interpreting,* Holly Mikkelson and Renée Jourdenais (eds), London/New York, Routledge: 352–67.

Braun, Sabine, and Judith L. Taylor (eds) (2012) *Videoconference and Remote Interpreting in Legal Proceedings*, Cambridge/Antwerp, Intersentia.

Braun, Sabine, and Judith L. Taylor (eds) (2014) *Advances in Videoconferencing and Interpreting in Legal Proceedings*, Cambridge/Antwerp, Intersentia.

Bührig, Kristin, and Bernd Meyer (2004) "*Ad hoc* Interpreting and Achievement of Communicative Purposes in Doctor-patient Communication" in *Multilingual communication*, Juliane House and Jochen Rehbein (eds), Amsterdam/Philadelphia, John Benjamins: 43–62.

Dal Fovo, Eugenia, and Natacha S. Niemants (eds) (2015) *Dialogue Interpreting*. Special issue, *The Interpreters' Newsletter* No. 20.

Davidson, Brad (2000) "The Interpreter as Institutional Gatekeeper: The Social-linguistic Role of Interpreters in Spanish-English Medical Discourse", *Journal of Sociolinguistics* 4, no. 3: 378–405.

Davidson, Brad (2002) "A Model for the Construction of Conversational Common Ground in Interpreted Discourse", *Journal of Pragmatics* 34, no. 9: 1273–300.

Davitti, Elena, and Sergio Pasquandrea (2014) "Guest Editorial", *The Interpreter and Translator Trainer* 8, no. 3: 329–35.

Falbo, Caterina (2012) "CorIT (Italian Television Interpreting Corpus): classification criteria" in *Breaking Ground in Corpus-based Interpreting Studies*, Francesco Straniero Sergio and Caterina Falbo (eds), Bern, Peter Lang: 155–85.

Gavioli, Laura (2012) "Minimal Responses in Interpreter-mediated Medical Talk" in *Coordinating Participation in Dialogue Interpreting,* Claudio Baraldi and Laura Gavioli (eds), Amsterdam/Philadelphia, John Benjamins: 201–28.

Gavioli, Laura (2015) "On the Distribution of Responsibilities in Treating Critical Issues in Interpreter-mediated Medical Consultations: The Case of 'le spieghi(amo)'", *Journal of Pragmatics* 76: 169–80.

Gavioli, Laura, Claudio Baraldi, and Natacha Niemants (2016) "From Archives to Corpora: Extracting Data for Analysis, from a Collection of Interpreter-mediated Interactions", contribution to the Panel "Working with corpora in community interpreting research: challenges and opportunities", Bernd Meyer (convenor), *8th International Critical Link conference*, 29 June 2016, Heriot Watt, Edinburgh, UK.

Hlvac, Jim (2013) "Should Interpreters be Trained and Tested in Telephone and Video-Link Interpreting? Responses from Practitioners and Examiners", *International Journal of Interpreter Education* 5, no. 1: 34–50.

Hsieh, Elaine (2007) "Interpreters as Co-diagnosticians: Overlapping Roles and Services between Providers and Interpreters", *Social Science & Medicine* 64, no. 4: 924–37.

Hsieh, Elaine (2010) "Provider-interpreter Collaboration in Bilingual Health Care: Competitions of Control over Interpreter-mediated Interactions", *Patient Education and Counselling* 78, no. 2: 154–9.

Jefferson, Gail (1972) "Side sequences" in *Studies in Social Interaction,* David N. Sudnow (ed), New York, NY, Free Press: 294–330.

Kelly, Nataly (2007) *Telephone Interpreting: A Comprehensive Guide to the Profession*, Bloomington, IN, Trafford Publishing.

Kelly, Nataly (2008*) A Medical Interpreter's Guide to Telephone Interpreting*, *International Medical Interpreters Association*, URL: **http://www.imiaweb.org/uploads/pages/380.pdf** (accessed 11 November 2016).

Ko, Leong (2006) "The Need for Long-term Empirical Studies in Remote Interpreting Research: A Case Study of Telephone Interpreting", *Linguistica Antverpiensia* 5: 325–38.

Lee, Jieun (2007) "Telephone Interpreting — Seen from the Interpreters' Perspective", *Interpreting* 9, no 2: 231–52.

Locatis, Craig, Williamson Deborah, Gould-Kabler Carrie, Zone-Smith Laurie, Detzler Isabel, Roberson Jason, Maisiak Richard, and Michael Ackerman (2010) "Comparing In-Person, Video, and Telephonic Medical Interpretation", *Journal of General Internal Medicine* 25, no. 4: 345–50.

Mason, Ian (eds) (1999) *Dialogue Interpreting*. Special Issue, *The Translator* 5, no. 2.

Mason, Ian (2006) "On Mutual Accessibility of Contextual Assumptions in Dialogue Interpreting", *Journal of Pragmatics* 38, no. 3: 359–73.

Merlini, Raffaela (2015) "Dialogue Interpreting", in *Routledge Encyclopedia of Interpreting Studies*, Franz Pöchhacker (ed.), London/New York, Routledge: 102–7.

Meyer, Bernd (2012) "*Ad hoc* Interpreting for Partially Language-proficient Patients: Participation in Multilingual Constellations" in *Coordinating Participation in Dialogue Interpreting,* Claudio Baraldi and Laura Gavioli (eds), Amsterdam/Philadelphia, John Benjamins: 99–113.

Moser-Mercer, Barbara (2003) "Remote Interpreting: Assessment of Human Factors and Performance Parameters", *aiic.net*. May 19, 2003. URL: **http://aiic.net/p/1125** (accessed 11 November 2016).

Moser-Mercer, Barbara (2005) “Remote Interpreting: Issues of Multi-sensory Integration in a Multilingual task”, *Meta* 50, no. 2: 727–38.

Niemants, Natacha (2012) “The Transcription of Interpreting Data”, *Interpreting* 14, no. 2: 165–91.

Niemants, Natacha (2015) *L’interprétation de dialogue en milieu médical. Du jeu de rôle à l'exercice d'une responsabilité*, Roma, Aracne.

Niemants, Natacha, and Sara Castagnoli (2015) “La traduction téléphonique en milieu médical : De l’analyse conversationnelle aux implications pratiques” in *Metamorfosi della traduzione, in ambito francese-italiano*, Danielle Londei, Sergio Poli, Anna Giaufret, e Micaela Rossi (eds), Genova, GUP: 227–262.

Price, Erika Leemann, Pérez-Stable Eliseo J., Nickleach Dana, López Monica, and Leah S. Karliner (2012) “Interpreter Perspectives of In-person, Telephonic, and Videoconferencing Medical Interpretation in Clinical Encounters”, *Patient Education and Counseling* 87, no. 2: 226–32.

Rosenberg, Brett Allen (2007) “**A Data Driven Analysis of Telephone Interpreting**” in *The Critical Link 4 – Professionalisation of Interpreting in the Community*, Cecilia Wadensjö, Birgitta Englund-Dimitrova, and Anna-Lena Nilsson (eds), Amsterdam/Philadelphia, John Benjamins: 65–75.

Roziner, Ilan and Miriam Shlesinger (2010) “Much Ado about Something Remote”, *Interpreting* 12, no. 2: 214–47.

Sacks, Harvey, Schegloff Emanuel A., and Gail Jefferson (1974) “A Simplest Systematics for the Organization of Turn-taking for Conversation”, *Language* 50, no. 4: 696–735.

Setton, Robin (2011) “Corpus-based Interpreting Studies (CIS): Overview and Prospects” in *Corpus-based Translation Studies – Research and Applications*, Alet Kruger, Kim Wallmach, and Jeremy Munday (eds), London, Continuum: 33–75.

Spinolo, Nicoletta, Bertozzi Michela, and Mariachiara Russo (forthcoming) “Shaping the interpreters of the future and of today: preliminary results of the SHIFT project”, *The Interpreters’ Newsletter* 23.

Straniero Sergio, Francesco, and Caterina Falbo (2012a) “Studying Interpreting through Corpora. An Introduction” in *Breaking Ground in Corpus-based Interpreting Studies*, Francesco Straniero Sergio and Caterina Falbo (eds), Bern, Peter Lang: 9–52.

Straniero Sergio, Francesco, and Caterina Falbo (eds) (2012b) *Breaking Ground in Corpus-based Interpreting Studies*. Bern: Peter Lang.

Villarruel, Antonia M., Portillo Carmen J., and Pamela Kane (1999) “Communicating with Limited English Proficiency Persons: Implications for Nursing Practice”, *Nursing Outlook* 47, no. 6: 262–70.

Wadensjö, Cecilia (1998) *Interpreting as Interaction*, London/New York, Longman.

Wadensjö, Cecilia (1999) “Telephone Interpreting and the Synchronisation of Talk in Social Interaction”, *The Translator* 5, no. 2: 247–64.

Zorzi, Daniela (2012) “Mediating Assessments in Healthcare Settings” in *Coordinating Participation in Dialogue Interpreting,* Claudio Baraldi and Laura Gavioli (eds), Amsterdam/Philadelphia: John Benjamins: 229–49.

Notes

Note: The authors have jointly discussed the contents of the paper, but primary responsibility for writing the different sections is as follows: Sara Castagnoli wrote sections 1, 3 and 5, while Natacha Niemants wrote sections 2 and 4.

[1] The SHIFT project (*SHaping the Interpreters of the Future and of Today,* **http://www.shiftinorality.eu**) is a 3-year Erasmus+ project funded by the European Commission in 2015 which aims to develop solutions for training in remote dialogue interpreting through the cooperation of a European network of universities offering interpreting programmes and interpreting service providers.

[2] AIM stands for Analysis of Interaction and Mediation and is the name of an Italian research network that has contributed to this collective project by sharing already transcribed data and/or by transcribing new subsets of audio-recorded interactions (**http://www.aim.unimore.it/**).

[3] Project title: *Analysis of communication with migrant patients and suggestions for improvements in the healthcare system* - P.I. Prof. Claudio Baraldi, University of Modena and Reggio Emilia, financed under the FAR 2014 competitive programme and concluded with an international seminar which took place in Modena on December 13, 2016.

[4] http://www.yorku.ca/comindat/comindat.htm

[5] Although these tag names are normally associated to written language corpora, they are temporarily used because of the specific requirements of the tools used to annotate and encode the corpus.

[6] Concordances in example (1) are taken from a version of the corpus that was encoded with the Corpus WorkBench (**http://cwb.sourceforge.net/**) and searched with the related Corpus Query Processor. Relevant annotations at the turn level are displayed in angle brackets before each actual concordance, where the search term is also enclosed in angle brackets and shown in boldface.

[7] Examples formatted as in (2) are taken directly from the xml version of the corpus.

"Corpora worth creating: A pilot study on telephone interpreting", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.

Stable URL: http://www.intralinea.org/specials/article/2315

"The translation challenges of premodified noun phrases in simultaneous interpreting from English into Italian", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2322

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy.

The translation challenges of premodified noun phrases in simultaneous interpreting from English into Italian

A corpus-based study on EPIC

By Serena Ghiselli (University of Bologna, Italy)

Abstract & Keywords

English:

This paper examines the handling of complex noun phrases in simultaneous interpretation into Italian of English speeches in the electronic corpus EPIC (European Parliament Interpreting Corpus). The complex noun phrases analysed in this study are noun phrases with two or more premodifying items included in the following categories: nouns, adjectives, adverbs, cardinal numbers and genitives. The aim is to extract complex noun phrases from a large sample of authentic English speeches and compare them with their corresponding translation into Italian in order to study the strategies used by interpreters. The initial hypothesis was that complex noun phrases pose a translation challenge in simultaneous interpreting from English into Italian because of structural and lexical diversities and memory overload. This hypothesis was partially confirmed in that strings where information was changed or deleted represent 45 per cent of the cases. In most cases, however, interpreters were able to adopt effective translation strategies.

Keywords: simultaneous interpreting, complex noun phrases, noun strings, corpus-based interpreting studies, strategies, premodifying items, european parliament

1. Introduction

The topic of this study is the translation of complex noun phrases in simultaneous interpreting from English into Italian. Complex noun phrases are common in English, a language in which attributive adjectives are placed before the noun they modify (Quirk et al. 1985: 402). In this study the term *complex noun phrases* indicates structures where the noun head is preceded by two or more modifiers placed next to one another or linked by the conjunctions *and, or* and *but*.

Handling complex noun phrases in simultaneous interpreting from English into Italian is a challenge as modifiers normally follow the noun head in Italian, thus taxing working memory. This issue has already been discussed in the graduation thesis of Barbafina (2003), who carried out an experiment with 10 interpreting students: they undertook a simultaneous interpreting exercise from English into Italian of a text to which long sequences of adjectives were added. Students tended to omit these sequences, especially when these did not change the overall meaning of the source text. Moreover, the high information density of the strings of modifiers brings about specific processing constraints due to memory overload. As Gile pointed out in his Effort Models' tightrope hypothesis (1997), interpreters have to coordinate various efforts and if one effort requires too many attentional resources interpreting performance may suffer from it.

Two studies tackled the same issue for two different language pairs: English into Hebrew and Polish into Italian. Shlesinger (2003) dealt with the memory overload in the translation of noun strings in simultaneous interpreting into Hebrew of English texts delivered at different speeds. Her hypothesis was that a slower presentation rate overloaded the interpreters' working memory more than a faster one. To test it, she carried out an experiment with 16 professional interpreters who interpreted six texts where strings of adjectives were added. Every text was interpreted twice (three weeks apart), once at 120 and the other at 140 wpm. The effect of presentation rate was in the predicted direction: a consistently better performance at a higher presentation rate, when there is less time for unrehearsed items to decay. Nevertheless, observed differences in performance were statistically non-significant. The other study is the doctoral dissertation of Cappelli (2014). Aiming to find out the strategies adopted by interpreters at the European Parliament to translate long strings of nouns from Polish into Italian, she observed that interpreters tend to omit some parts that can be inferred from the context.

The present paper is based on the author's graduation thesis (Ghiselli 2015), which was inspired by the widespread perception among interpreting students that the translation of complex noun strings from English into Italian requires increased cognitive effort. This study is corpus-based and includes the analysis of complex noun phrases extracted from original English speeches contained in an electronic corpus and their simultaneous interpretations into Italian. The corpus is the *European Parliament Interpreting Corpus* (EPIC) (Sandrelli et al. 2010; Russo et al. 2012). The available metadata are specified in the transcript header of every text in the corpus. These are a sequence of fields providing information about the speaker (gender, country, mother tongue, political function and group) and about the speech (date, id number, language, type, duration, timing, text length, number of words, speed, words per minute, source text delivery, topic, specific topic).

The aims of the research were, firstly, to identify the strategies used by EU interpreters to translate complex noun strings and, secondly, to characterise the extent to which these strategies were influenced by speed and mode of delivery. In the EPIC corpus the speed of delivery can be *low* (< 130 w/m), *medium* (130–160 w/m) and *high* (> 160 wpm). The mode of delivery of the text is labelled *impromptu, read* or *mixed* (Monti et al. 2005). The hypothesis is that the translation of complex noun phrases from English into Italian in simultaneous interpreting is difficult to handle in texts delivered quickly and texts read from written notes. Texts delivered quickly present greater difficulties because of higher cognitive load for interpreters due to time constraints, whilst texts read from written

notes have a higher number of noun phrases. Some characteristics of the source text, such as redundancy, familiarity or explicitness, have an impact on the level of difficulty in interpreting it simultaneously (Hönig 2002; Alexieva 1994, 1999). Alexieva (1999) demonstrated that texts with more than one (two, three or even more) implicit predications are difficult to comprehend. She collected data from different sources such as four interpreting classes (50 students), summary writing exercises (60 students), multiple choice listening comprehension tests (65 trainees) and answers elicited from interpreters used as informants. Moreover, almost all the mistakes in interpreting were found in the highly condensed portions of the text, or the parts after them. As far as the mode of delivery is concerned, Hönig (2002) points out that a speech delivered using a prepared manuscript is more difficult than an impromptu speech because the speaker might read it at a very high speed, giving the wrong emphasis and occasionally leaving out words.

The paper will start by describing the materials of the study; then the way in which noun phrases were extracted from the corpus and divided into different categories will be outlined; finally, there will be the discussion of results and some conclusive remarks.

2. Materials and Methods

2.1 Materials

2.1.1 The corpus

The source of the speeches used for this study is the European Parliament Interpreting Corpus (EPIC) (Monti et al. 2005; Sandrelli, Bendazzoli and Russo 2010; Russo et al. 2012). In corpus linguistics a corpus is a large collection of authentic texts in electronic format created according to a set of criteria and is characterised by representativeness, dimension and format (Bowker and Pearson 2002: 9). A very interesting aspect of electronic corpora is the opportunity to perform *semi-automatic searches* thanks to *markup*. Markup is related to the description and explicitation of the structure of a certain text. *Tagging* is related to a more specific level in the texts, that is linguistic and pragmatic aspects (Bendazzoli 2010: 76). *POS-tagging* in particular is a considerable added value for a corpus because it makes it possible to do an automatic search of words in specific linguistic structures (Monti et al. 2005).

EPIC contains transcripts of speeches delivered at the European Parliament in February 2004 in English, Italian and Spanish and the interpretation of each speech into the two other languages involved. It was created by the research group of corpus-based interpreting studies of the then Department of Interdisciplinary Studies in Translation, Languages and Cultures (SITLeC) of the University of Bologna (now Department of Interpreting and Translation) with the aim of studying interpreting strategies and problems related to language pairs. EPIC is structured into *nine sub-corpora*. The comparable analysis carried out for this study is based on two sub-corpora: the sub-corpus of English source speeches (SS) and the corresponding sub-corpus of Italian target speeches (TS). The SS analysed are 81, for a total of 42,705 words.

Part of the EPIC archive was transcribed and is accessible online through the SSLMITDev website. The corpus is tagged, so texts can be automatically retrieved. The taggers used are *Tree Tagger* for Italian and English and *Freeling* for Spanish. They were created for written texts and, therefore, some tags were wrongly assigned in the oral speeches of EPIC. However, generally speaking, tagging was satisfactory, with more than 90% of correct tags in all subcorpora (Bendazzoli 2010: 131). EPIC can be queried with *simple or advanced queries*. Simple queries can be used to find words or strings of words in context, whereas advanced queries have to be written in the *Corpus Query Processor (CQP)* language and aim at finding occurrences of complex strings of elements (Bendazzoli 2010: 134–35).

2.1.2 Noun phrase modifiers

In English noun phrases consist of a head, normally a noun, and of elements that determine and optionally modify the head or complement another element in the phrase, for example *all those fine warm days in the country last year* (Quirk et al. 1985: 62). The complex noun phrases analysed here are noun phrases with two or more premodifying items from the following categories: adjectives, cardinal numbers, adverbs, nouns and genitives.

All the adjectives included in this study have an attributive function, meaning that they occur between the noun they modify and the determiner, as in *an ugly painting* (Quirk et al. 1985: 402–3). The intensifier *very* and the premodifiers *more* and *most* can have both the function of determiners and of adjectives (Collins English Dictionary, **http://www.collinsdictionary.com/dictionary/english/more?showCookiePolicy=true** and **http://www.collinsdictionary.com/dictionary/english/most?showCookiePolicy=true**, 21 April 2016) and have been counted as part of noun strings. Many adjectives in English have the same suffixes as participles in *–ing* or *–ed* and are called *participial adjectives,* for example: *his surprising views; the offended man* (Quirk et al. 1985: 413). Gerunds and present participles have been included in the study because, in the attributive function, they qualify the head of the noun phrase. In the noun strings of this study, adverbs having the function of adjective modifiers are included, as in *very vibrant poultry industry* (org-en text 4). Normally adverbs that premodify an adjective have the function of *intensifiers* and are used with an adjective having comparative and superlative forms (Quirk et al. 1985: 445). Genitives are most often used for possessions, relationships and physical characteristics, especially when the first noun refers to a person or animal, or to a country, organisation or other group of living creatures (Swan 2005: 440). Descriptive genitive (for example: *a women's college= a college for women*) acts as modifier and has a *classifying role* similar to that of noun modifiers and some adjective modifiers (Quirk et al. 1985: 327).

In Italian, the adjectival phrase can have different functions. It has an *attributive function* when it is in a noun phrase before or after the name that defines the phrase. Adjectives can be in a prenominal or in a postnominal position: in the noun phrase the adjective follows the name in the unmarked case and it precedes it in the marked one (Renzi et al. 2001: 439–40). An adverb can modify an adjective and it normally precedes it. In Italian the noun has inflection for gender (masculine/feminine) and number (singular/plural). Nouns are the grammatical heads of the noun phrase that influence the gender and number agreement of the other elements and of the predicate.

If the position of adjectives in respect of the noun is taken into account, they behave similarly in English and in Italian (Rosato 2013: 31). Compared to English, Italian has a "mirror image" post-nominal ordering: in English the noun head is at the end, whereas in Italian it is at the beginning of the noun phrase. For example, *dry red wine* becomes *vino rosso secco*. If we consider the distance of adjectives from the noun, we can notice that the noun head has a different position but the adjective order is the same.

2.2 Methods

2.2.1 Extraction of noun phrases from EPIC

The first phase of this corpus study consisted of elaborating search queries from the *Advanced query* interface in EPIC in order to find complex noun phrases. As already stated, EPIC is POS-tagged, so it was possible to perform not only searches by single words but also searches by parts of speech, it was therefore possible to perform an automatic search and then manually select results following inclusion and exclusion criteria. The text numbers mentioned in this article are the *Text ID* numbers that appear on top of the page when opening a search result in SSLMITDev. Every original speech shares the same *Text ID* number with its corresponding interpretation into Italian.

The subcorpus of original English speeches (org-en) was queried by means of three different advanced query expressions. The search parameters were: *find at maximum 10,000 results* (the highest possible option). The *Results set* was *Random set,* that is the visualisation of all the possible results and the *Results per page* option was *no limit.* The aim of the three expressions was to cover all possible kinds of premodified noun phrases.

The first expression, which will be called *expression A*, recalled a noun preceded by at least two modifiers, for example *larger commercial units* (org-en text 1). The tags of all possible modifiers were included in the first two parts of the expression, that is the tags for adjectives, cardinal numbers, nouns, past participle, *-ing* forms, possessive pronouns and adverbs. For the head of the noun phrase all the tags for nouns were used. *Two modifiers + noun* was set as the minimum required length for the strings to be included in the study, but there are also longer strings, which could be identified through the *key word in context (KWIC)* consultation of the corpus. The *context* was of *25 characters* before and after the result.

The second expression, which will be called *expression B*, looked for a noun preceded by two modifiers linked by a coordinative conjunction such as *and*, *or, but*, for example *Food and Veterinary Office* (org-en text 7). Premodifiers were searched using the same tags as in *expression A*, the coordinative conjunctions were retrieved through the tag of coordinating conjunctions and, for the noun head, all the tags for nouns were included.

The third expression, which will be called *expression C*, consisted of tags for possessive endings. When this search was performed, it only provided 146 occurrences. Nonetheless, it was decided not to change it by adding further search parameters and results were manually selected. Genitives were taken into consideration only when they were premodifiers of a noun phrase with at least another premodifier, for example *last year's second preparatory committee* (org-en text 27).

An automatic system of data collection was chosen because it was quicker and reliable in terms of including all the potentially relevant results. However, it was just a first phase of the research because while reading the results it was clear that the *Advanced query* function was not able to assess whether noun sequences made sense or not. For this reason, the automatic data extraction from EPIC was followed by a manual selection of results.

To ensure that the manual phase was as objective as possible, a set of inclusion and exclusion criteria for modifiers was drawn up. Adjectives, numbers, nouns, possessives and adverbs were looked up. Determiners (except for possessives) and predeterminers were excluded, as well as results which were considered irrelevant for different reasons, detailed in paragraph 2.2.3.

2.2.2 Inclusion criteria

The modifiers included in the search were all adjectives, cardinal numbers, nouns, past participles and gerunds, possessives and adverbs. For the noun head, all nouns were looked up.

Tags were normally correct, but there were some exceptions: *-ing* and *-ed* forms are always tagged as *VVG (verb gerund/participle)* and *VVN (verb past participle)* respectively. However, these forms are not always verbs, they can also be adjectives, as for instance in: *mounting circumstantial evidence* (org-en text 7) and *internal organised crime* (org-en text 18). In both cases the *-ing* and *-ed* forms are used as adjectives, but they would not have been retrieved from the corpus without the inclusion of the tags *VVG* and *VVN* in the search string.

Compound nouns of countries such as *United States*, or of organisations like *European Union*, were included in search results and treated as two-element items.

There are three kinds of peculiar noun phrases that have been included in the study. The first type includes eight strings retrieved with *expression A* that are more complex versions of *expression B*. An example is the following string from org-en text 7: *its sanitary and its economical dimension.* In this string there is a first modifier (*its sanitary*) + a conjunction (*and*) + a second modifier (*its economical*) + a noun head (*dimension*).

The second type is the case of three complex strings having multiple heads with the same modifiers. Multiple heads with the same modifiers are noun phrases where the same modifier applies to two nouns (Quirk et al. 1985: 1345–46). If there is just one modifier, the structure is not complex and is not in the remit of this study. Using the KWIC visualisation, two results of *expression A* and one result of *expression B* were identified as multiple heads with multiple shared modifiers, to be included in the study. These three strings are: *global and regional peace and security* (org-en text 27), *largest drug dealers and drug producers* (org-en text 37), *clear agreed objectives and positions* (org-en text 75).

The third and last special type of strings is the case of constructions with the preposition *of* used as premodifiers. There are two strings in org-en text 18 and one string in org-en text 28 where a prepositional phrase with *of* is one of the noun modifiers. These two complex constructions were retrieved with *expression A* and included in the results. These three strings are: *nineteen ninety six Hague Protection ehm of Children Convention*, *Chief of Police Task Force*, *Cold War weapons of mass destruction programmes*.

2.2.3 Exclusion criteria

Some exclusion criteria were applied before the string analysis was undertaken, whereas other criteria were developed during data collection. Determiners and predeterminers were excluded from the relevant modifiers because including them would have resulted in a much higher number of hits, the majority of which would not have been relevant. Determiners are elements at the beginning of noun phrases different from adjectives, for example *a, this, some, either, every, enough, several* (Swan 2005: 154). The only determiners that were included in the study are possessives. Prepositional structures with *of* were excluded, except for those mentioned before (see 2.2.2).

During data collection 14 exclusion criteria were developed:

1. Strings including *-ed* forms having a verbal function: the *-ed* forms that were excluded from the results were past participles preceded by the auxiliary verb *be* to form either the present perfect or the passive or implicit sentences including past participle.

2. Strings including *-ing* forms having a verbal function: *-ing* forms were excluded when they were present participles, gerunds or when they were used after prepositions.
3. Strings including *one's own + noun*: even if *own* is tagged as an adjective, it is solely an intensifier and adds no new information. Strings with *one's own* were included only if, apart from *one's own*, they included also other modifiers, for example: *my own little hairdresser* (org-en text 22).
4. Strings with numbers: in EPIC all the numbers are expressed in words, so they appeared in the results as word strings but were excluded (with the exception of numbers modifying a noun together with other modifiers) because they were not relevant for the study purpose.
5. Strings including time: for example *twelve o'clock* (org-en text 7).
6. Strings with lists: lists are formed by independent words and can be repeated in the same order, so they are not relevant for the present study.
7. Strings with vocatives: vocatives are expressions that speakers use to mention somebody among the public. They are very common at the beginning of EPIC speeches, when the speaker thanks the President of the Parliament or other subjects.
8. Strings including filled pauses: *Tree Tagger* often tags filled pauses as nouns. For this reason, the automatic search extracted a number of irrelevant results, which were excluded manually.
9. Long strings that appear in more than one result: phrases longer than the three-element first search string were retrieved more than once in groups of three elements. For these cases, the first occurrence was counted and the following excluded.
10. Strings containing wrong tags: in some cases results were not relevant because tags were wrong. The corpus tagger often made mistakes when there were truncated words, for example: *parliamentary secretariats a-(and)* (org-en text 12).
11. Strings with untranslatable proper names: these names had just to be repeated by the interpreters and were not relevant for the present study.
12. Strings with hesitations: strings containing repetitions, truncated words and filled pauses were excluded because they were not complex noun phrases. For example: *Lisbon pro- pro- process* (org-en text 71).
13. Strings with rephrasing: cases when the speaker rephrases the sentence, often with the aim of giving a more precise information. For example: *humans nineteen humans* (org-en text 12).
14. Special cases: strings that were excluded but which do not fit into any of the previous criteria. Many of these excluded results include genitives that are not part of a complex noun phrase, such as *girls' schools* (org-en text 33). There are also occurrences of adjacent nouns belonging to different phrases or without mutual relationship. In the KWIC visualisation they appear as in the following example: *situation of the European economy and major guidelines for economic policy and* (org-en text 51).

2.2.4 Noun phrases in relation to speed and mode of delivery

In the EPIC corpus transcript headers contain some information about the speaker and the text, among which two variables were considered interesting for data analysis: speed of delivery and mode of delivery. Text speed was calculated in words per minute and the texts of the corpus are divided into three types according to their speed: *low speed* (< 130 w/m), *medium speed* (130–160 w/m) and *high speed* (> 160 wpm). The mode of delivery of the text was classified as either *impromptu, read* or *mixed* (Monti et al. 2005).

Data about noun phrases were matched with speed and mode of delivery. The relation between speed and mode of delivery and the percentage of string words as a share of the overall number of words in every original English text was calculated.

The differential use of translation strategies found in target texts was analysed according to speed and mode of delivery. This analysis was based on two data sets created on the basis of Schjoldager's theoretical model of translation relationships (Schjoldager 1995, see section 2.2.5 below). The first data set cross-classifies Schjoldager's categories with delivery mode and delivery rate. The second data set includes the percentages of strings of every Schjoldager's category compared to the overall number of strings only in the texts having the same speed or the same mode of delivery.

2.2.5 Translational norms in simultaneous interpreting: Schjoldager's categorisation

In order to assess the translation strategies used by interpreters, it was necessary to divide them into categories and the categorisation used was Schjoldager's theoretical model of translation relationships (1995).

Schjoldager explores the potential of Toury's (1980) translational norms in the field of simultaneous interpreting studies. Schjoldager speculates whether Toury's norms can be applied to simultaneous interpreting and to what extent the cognitive complexity of this activity influences the use of translation strategies by interpreters.

Schjoldager identified five main categories, one of which is divided into six subgroups.

A/ Repetition: target-text item bears formal relation with relevant source-text item. Examples:

> *Eighteen human lives – diciotto vite umane [eighteen human lives]* (text 1)
>
> *Public health issues – temi di salute pubblica [issues of public health]* (text 7)
>
> *World Health Organisation – Organizzazione Mondiale per la Salute [World organization for health]* (text 11)
>
> *Regulatory and supervisory regime – regime normativo e di sorveglianza [regulatory regime and of supervision]* (text 43)
>
> *Increasingly intensive cooperation – cooperazione sempre più intensa [more and more intense cooperation]* (text 76)

B/ Permutation: target-text item(s) is(are) placed in a different textual position from relevant source-text item(s). Examples:

Pandemic influenza preparedness – preparazione alla pandemia e all'influenza [preparation to pandemic and to influenza] (text 1)

Member State responsibility – lo Stato membro responsabile [the accountable Member State] (text 18)

Its fundamentalist Islamic revolution - integralismo e la rivoluzione isl- islamica [extremism and the islamic revolution] (text 46)

European Union economy – Unione europea e la sua economia [European Union and its economy] (text 65)

Full and unconditional cooperation – piena ehm co- collaborazione senza condizioni [full ehm co-cooperation without conditions] (text 81)

C/ Addition: target-text item constitutes an addition to information given in relevant source-text item. Examples:

Animal health pro- issue – questione anche anche di salute animale [issue also also of animal health] (text 2)

Member States' responsibilities – responsabilità dei vari Stati membri [responsibility of the various Member States] (text 11)

Their long-term economic development - loro sviluppo economico a più lungo termine [their economic development at a longer term] (text 40)

Most competitive economy – principale economia e più competitiva [main and more competitive economy] (text 68)

Considerable simbolic importance - significato simbolico importante notevole [important considerable simbolic significance] (text 74)

D/ Deletion: no target-text item bears direct relation with relevant source-text item.

E/ Substitution: target-text item bears no formal relation with relevant source-text item.

E1/ Equivalent Substitution: source-text item is translated functionally. Example: *Larger commercial units – grandi aziende [large businesses]* (text 1); *Rome two regulations – regolamento Roma due [Rome two regulation]* (text 18); *my last remark – quest'ultima osservazione [this last remark]* (text 25).

E2/ Paraphrastic Substitution: source-text item is translated functionally, but in an expanded and/or segmental way. Example: *negative ehm criminal issue - non è tutto negativo non si parla solo di crimini [not everything is negative, you do not only talk about crime]* (text 22); *new and challenging agenda - ordine del giorno che sia una vera ehm sfida [agenda that would be a real ehm challenge]* (text 24)*; newly independent states – nuovi paesi indipendenti nuovi stati indipendenti [new independent countries new independent states]* (text 75).

E3/ Specifying Substitution: source-text item is translated functionally and implicit information is made explicit. *Example: military and security dimensions – dimensione della sicurezza e la soluzione del conflitto [security dimension and conflict resolution]* (text 16); *last few years – ultimo due anni [last two years]* (text 21); *our neighbourhood policy – nostra politica di buon vicinato [our policy of good neighbourhood]* (text 75).

E4/Generalizing Substitution: source-text item is translated functionally, but conveys less information than relevant source-text item. *Example*: *very worrying aspect – aspetto preoccupante [worrying aspect]* (text 4); *their hard work – lavoro [work]* (text 11); *important Additional Protocol – protocollo aggiuntivo [additional protocol]* (text 28).

E5/ Overlapping Substitution: source-text item is translated functionally, but with a different viewpoint, so that target-text item conveys different information. Example: *FAO-WHO-OIE expert panel – gruppo di esperti di alto livello* (text 1); *two further amendments – due blocchi di emendamenti [two blocks of amendments]* (text 12); *very little progress – qualche progresso [some progress]* (text 30).

E6/ Substitution Proper: target-text item bears little or no resemblance to relevant source-text item. Example: *only ninety inspectors – sforzi enormi da questi ispettori [enormous efforts from those inspectors]* (text 2); *clear scientific leadership – dati scientifici [scientific data]* (text 11); *richest and most prosperous parts – guardate quello che è successo [look at what happened]* (text 62).

The translations of the selected strings were found by comparing the original English text with the corresponding interpretation into Italian using the author's knowledge of both languages. The noun strings were matched this way with their translations and then classified according to Schjoldager's categories.

An effort was made to keep the division of the strings' translations into the different categories as objective as possible. However, some categories cannot always be clearly distinguished and differences of interpretation may occur. Thus, the categories *E1/ Equivalent Substitution*, *E4/ Generalizing Substitution* and *E5/ Overlapping Substitution* contained some items which could, with justification, have been reassigned. A consistent approach was adopted throughout the whole corpus. For example, all the strings where every word of the translation corresponded to the original text with a minor change concerning a singular turned into a plural or vice versa were put into *E1/ Equivalent Substitution*.

3. Results and discussion

3.1 Results

The subcorpus org-en has 42,705 words and 970 strings were retrieved in the study. This corresponds to 3482 words (7.9 % of the words in the subcorpus). The average string length was 3.59 words. The classification of the strings' translations according to Schjoldager's categories is shown in Table 1 and Figure 1.

Strings divided by translation category

Translation categories	Number of strings	% of total number of strings
A/ Repetition	343	35.36
B/ Permutation	27	2.78

C/ Addition	16	1.65
D/ Deletion	98	10.10
E1/ Equivalent Substitution	93	9.59
E2/ Paraphrastic Substitution	40	4.12
E3/ Specifying Substitution	12	1.24
E4/ Generalizing Substitution	265	27.32
E5/ Overlapping Substitution	40	4.12
E6/ Substitution Proper	36	3.71
Overall number of strings	**970**	100.00

Table 1: Strings by translation category

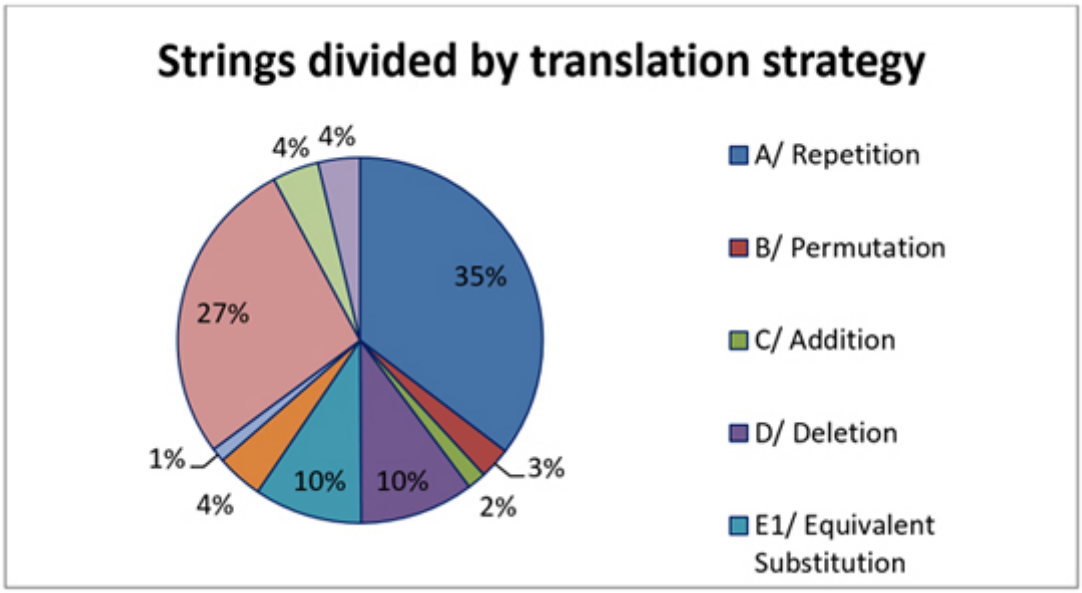


Figure 1: Strings per translation strategy

Schjoldager's categories were divided into two groups. The first group consists of strategies that give evidence of interpreters' control: *A/Repetition, B/Permutation, C/Addition, E1/Equivalent Substitution, E2/Paraphrastic Substitution* and *E3/Specifying Substitution*. In the strings belonging to these categories no elements were lost, the elements of the noun phrase appeared in the same order or in a different order as the original text (A and B), they were reformulated (E1 and E2) or even completed and expanded (C and E3). The second group of categories includes problematic strategies, including *D/Deletion, E4/Generalizing Substitution, E5/Overlapping Substitution* and *E6/Substitution Proper*. In these categories, strings are modified: the original message was completely or partially lost (D and E4) or partially modified or replaced with different elements (E5 and E6).

Category E4 was then analysed more in detail, leading to the identification of two further subcategories: *subcategory one* (strings with modifiers omitted) and *subcategory two* (strings with noun heads omitted). Strings of subcategory one were then divided into four groups: strings with two, three, four or five premodifiers. It appeared that 90 per cent of cases belong to subcategory one and only 10 per cent to subcategory two. The percentage of omissions of modifiers is similar in the four groups and equal to 57 per cent.

3.2 Discussion

The hypothesis was that read texts were denser and therefore had more complex noun phrases, an issue that had already been considered by Hönig (2002) and Alexieva (1994, 1999). In this study, complex noun phrases are strings where the noun head is preceded by two or more modifiers placed next to one another or linked by the conjunctions *and, or* and *but*.

The most used translation strategy was *A/Repetition* (see Chart 1), which is noteworthy, as it is a successful outcome, because it shows that the interpreter has fully understood the message in the source language and then correctly and exhaustively reproduced it in the target language. In general, 55 per cent of the results belong to one of the successful strategies (categories A, B, C, E1, E2 and E3)*;* the remaining 45 per cent displayed a more problematic strategy (categories D, E4, E5 and E6). It can thus be concluded that interpreters at the European Parliament handle complex noun phrases rather well.

Some omissions in category *E4/ Generalizing Substitution* were associated with a substantial loss of information. A recurrent strategy to compensate for this loss is replacement of premodifiers with more general expressions such as indefinite or demonstrative adjectives, for example: *two ongoing EC assistance programmes – altri programmi [other programs]* (text 1) and *its traditionally long historic record – questo questa lunga storia di progresso [this this long record of progress]* (text 46). Other omissions do not imply a substantial loss of information instead: 36 premodifiers omitted in strings belonging to category E/4 are possessives and, among them, in 30 cases the possessive is the only premodifier omitted. In some cases possessives in Italian would have been redundant, as in the following example from text 66: *our common prosperity – prosperità comune [common prosperity]*. Another trend is the omission of references to Europe, like the acronym *EU,* the nouns *Europe and European Union* and the adjective *European*. Taking the context into account, the plenary sitting of the European Parliament, those references

are often already clear to the listener. The strategy of omitting context-deducible elements confirms the results both of Cappelli's PhD thesis (Cappelli 2014) and of Barbafina's graduation thesis (Barbafina 2003). As mentioned in the introduction, Cappelli's study dealt with the interpretation of complex noun strings in a different language pair, namely Polish into Italian. She analysed interpreted speeches at the European Parliament and found that in 41 per cent of cases some elements of the noun phrases were missing, but these omissions could be inferred by the context. Barbafina carried out a study about the simultaneous interpreting of long sequences of adjectives in the same language pair considered in the present study, that is English into Italian, with advanced students of interpreting at the University of Bologna. Students tended to omit problematic items, especially when they did not change the overall meaning of the source text, and few sequences were translated completely and correctly. Thus, Barbafina's results support the assumption upon which her study and the study described in this paper are based: long sequences of elements indeed represent a potential problem for interpreters in simultaneous interpreting. In EPIC, in contrast to Barbafina's study, there are some omissions but the majority of strings are translated completely and correctly.

It may be hypothesised that interpretation of complex noun phrases improves with the acquisition of strategies via experience. This might be an interesting research question for a comparable study adopting the novice versus expert paradigm. Applying strategies is in fact crucial in simultaneous interpreting and there are several studies on this topic. Sunnari (1995), for instance, argues that interpreting exhaustively may not always be the best choice. Expertise means knowing what can be left out without losing key ideas of the text. Eliminating redundancy, for example, is a strategic choice showing expertise. By applying certain mental operations (macrorules) to the source language message (microstructure) during comprehension, interpreters should be constructing the macrostructure of what they hear. Riccardi (2005) distinguishes between skill-based and knowledge-based strategies in simultaneous interpreting. Skill-based strategies are governed by stored patterns of automatic responses whose application is triggered by the recognition of a well-known stimulus within the communicative event. Knowledge-based strategies are conscious and come into play when actions must be planned on-line because no automatic response is found or because something has caused a momentary memory overload. Bartlomiejczyk (2006) carried out an experiment about interpreting strategies and directionality with English and Polish, including retrospective remarks of the interpreters, and she identified 21 interpreting strategies. Liu (2008) describes expertise in simultaneous interpreting as the result of well-practised strategies in each of the comprehension, translation and production processes and the interaction among these processes, which are specific to the needs of the task of simultaneous interpreting. Kader and Seubert (2015) distinguish between macro-strategies, that include planning and expectations before the assignment, and micro-strategies, which are related to speech-inherent issues.

3.2.1 Strings in relation to speed and mode of delivery

The data presented in this study confirm the hypothesis that read texts have more complex noun phrases. A further important observation is that impromptu speeches have a much lower percentage of strings found in speeches delivered in both read and mixed modes. On average, complex noun phrases in terms of word count represent 7.9 per cent of the text in the subcorpus org-en. The percentages for each subgroup of mode of delivery, by contrast, were 4.18 per cent (impromptu), 9.52 per cent (read) and 8.23 per cent (mixed).

Speeches delivered at low speed are handled well by interpreters: 60.87 per cent of performances belong to categories of successful interpretation. Complete omissions accounted for 2.48 per cent of all source strings and the most used strategy was *A/Repetition* (42.86 per cent). Moreover, *B/Permutation* was observed in 29.63 per cent of the cases in these texts. Interpreters thus seemed to have enough time both to translate and to reorganise the phrase elements in a different order. In speeches delivered at medium speed, 58.89 per cent of the occurrences belonged to the categories of successful interpretation. *A/Repetition* (35.45 per cent) is the most frequently used strategy, but the percentage of *D/Deletion* was three times the rate found in texts delivered at a low speed (8.97 per cent). *A/Repetition* was also the most used strategy in texts delivered at high speed (31.53 per cent). Source texts with high delivery rates are the only group for which the percentage of problematic strategies (51.43 per cent) is higher than that of successful strategies (48.57 per cent). Moreover, this group had the highest percentage of *D/Deletion* (15.06 per cent). It is noteworthy that more than half of the occurrences of *D/Deletion* (54.08 per cent) and of *E6/ Substitution Proper* (55.56 per cent) were found in texts delivered at high speed.

In summary, the percentage of *A/Repetition* decreases as delivery speed increases, whereas the percentage of *D/Deletion* of strings increases as delivery speed increases. The initial hypothesis that complex noun phrases are more difficult to handle in fast speeches was confirmed by the study data. This finding contrasts with the findings of Shlesinger (2003) in her study of memory overload in the translation of noun strings in simultaneous interpreting from English into Hebrew. Shlesinger observed important, even if statistically non-significant, differences in performance, with more modifiers retained in texts presented at a higher speed. Possible explanations of the discrepancy may be that Shlesinger's analysis is based on experimental data and that it focuses on adjectival modifiers, reporting on retention of those modifiers only, without considering the omissions of the noun head. When Shlesinger talks about the materials she used, she points out that, in terms of ecological validity, the materials for such a study should ideally be taken from an existing corpus of actual conference presentations. This was not possible in her case, because her study required clearly delineated, accurately constructed strings, unlikely to occur in naturalistic speech. For the present study, it was possible not only to consider adjectival modifiers but also to access EPIC, a validated corpus-based resource of real speeches. Moreover, different parameters for speed were used in the two studies. Speed of delivery at the European Parliament is on average higher than what is found at a conference because speakers are allotted very short time slots to deliver their speeches. In Shlesinger's experiment, the high delivery rate was set at 140 wpm, whereas in EPIC original English speeches the average delivery rate of high speed texts is 180 wpm. It can be argued that, if the delivery rate is so high, other difficulties add to memory overload, for example a greater effort to coordinate listening and speaking, leading to a very challenging translation of complex noun phrases.

4. Conclusions

The initial hypothesis that complex noun phrases in English are a challenge for interpreters translating into Italian was only partially confirmed in the data analysis from EPIC. If we consider the whole subcorpus of original English speeches, the percentage of strings belonging to Schjoldager's problematic categories is 45, less than half of the occurrences. On the other hand, 45 per cent is quite high, so it can be said that, even if interpreters at the European Parliament handle complex noun phrases well in the majority of cases, occurrences of incomplete or wrong translations are also rather common. The most common translation strategies are complete and correct translations and generalisations, with 35 per cent and 27 per cent of the occurrences respectively. Less frequent strategies are

additions (1.6 per cent) and specifications (1.2 per cent): this is unsurprising as complex noun phrases are already dense and adding more information is often not feasible.

Therefore, the translation of complex noun phrases is challenging and special attention in training and practice of simultaneous interpreting is recommended. For example, the control of Ear-Voice-Span (EVS) is a crucial issue. Van Dam (1989) describes a set of beginner exercises which include distance exercises to avoid tailgating the speaker and keeping the optimum start-up distance, which corresponds to approximately one meaning unit behind the speaker at the beginning of each new sentence. EVS has to vary during the task and a balance should be struck between understanding the message before speaking and not overloading working memory.

Kader and Seubert (2015) include flexible EVS in micro-strategies and write that complex passages, such as lists, demand a short EVS, whereas more abstract ideas require a longer one. According to Liu (2008), experts use semantic-based processing strategies to free up mental resources. In this way the interpreter is able to anticipate the upcoming information based on the context that is provided. One semantic processing strategy is expert interpreters' ability to perceive and distinguish the importance of the input material and to pay more attention to the overall conceptual framework of the source speech, which may also contribute to their ability to segment the input material into bigger chunks during the process of translation.

Among Bartlomiejczyk's (2006) strategies, compression might be useful in the case of complex noun phrases. Compression means summarising a longer fragment with a shorter phrase, which is supposed to convey the same meaning, but expressed in a more concise and general way. This strategy is similar to Kader and Seubert' condensing strategy (2015). According to Riccardi (2005), with increasing expertise the primary focus of control moves from the knowledge-based (conscious) to the skill-based (automatic) level, providing for a well-balanced allocation of cognitive resources.

Translating complex noun phrases in simultaneous interpreting from English into Italian is a challenge for many interpreters, but only limited research has been previously undertaken on this topic. This study is offered as a contribution based on real data coming from a specific setting, the European Parliament. The study includes various types of modifiers and focuses on the number of omissions and on retentions and their translation. Another aspect that has not been investigated and that could be interesting for a more fine-grained approach is whether there are differences between the different types of modifiers, for instance whether the adjectival modifiers are more vulnerable to deletion than the nominal ones. Another further development is analysing what exactly is omitted and to what extent. It stands to reason, for instance, that the omission of the head noun is much more harming to the general understanding than the omission of a modifier. Research with more data and in other domains is needed to develop a better understanding of the difficulties posed by noun strings. Further research questions may also include the difference between the novice and the expert interpreters' approach and whether a high-level of expertise in a specific field, acquired thanks to specialised studies and experience in that sector, has a positive impact on the handling by interpreters of the complex noun phrase terms belonging to that field. Skill-based strategies seem to be the hallmarks of expertise and flexible EVS makes it possible to keep the optimum start-up distance and adopt more semantic-based processing strategies. Starting from this consideration, a hypothesis for further research may be that students benefit from automatising as many linguistic expressions as they can and should focus on improving working memory, so that they will have more cognitive resources to keep an EVS that allows them to make the more appropriate strategic choice when translating complex noun phrases.

References

Alexieva, Bistra (1994) "Types of Texts and Intertextuality in Simultaneous Interpreting" in *Translation Studies: An Interdiscipline*, Mary Snell Hornby, Franz Pöchhacker, Klaus Kaindl (eds), Amsterdam, John Benjamins: 179–88.

Alexieva, Bistra (1999) "Understanding the Source Language Text in Simultaneous Interpreting", *The Interpreters' Newsletter* 9: 45–59.

Bartlomiejczyk, Magdalena (2006) "Strategies of Simultaneous Interpreting and Directionality", *Interpreting* 8, no. 2: 149–74.

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

Bowker, Lynne and Jennifer Pearson (2002) *Working with Specialised Language – A Practical Guide to Using Corpora*, London, Routledge.

Gile, Daniel (1997) "Conference Interpreting as a Cognitive Management Problem" in *Cognitive Processes in Translation and Interpreting*, Joseph H. Danks, Gregory M. Shreve, Stephen B. Fountain, and Michael McBeath (eds), Thousand Oaks, London, Sage Publications: 196–214.

Hönig, Hans G. (2002) "Piece of Cake - or Hard to Take? Objective Grades of Difficulty of Speeches Used in Interpreting Training" in *Teaching Simultaneous Interpretation into a "B" Language*, EMCI Workshop, 20–21 September 2002.

Kader, Stephanie, and Sabine Seubert (2015) "Anticipation, Segmentation…Stalling? How to Teach Interpreting Strategies" in *To Know How To Suggest… Approaches to Teaching Conference Interpreting,* Dörte Andres, Martina Behr (eds), Berlin, Frank and Timme: 125–44.

Liu, Minhua (2009) "How do Experts Interpret? Implications from Research in Interpreting Studies and Cognitive Science" in *Efforts and Models in Interpreting and Translation Research. A Tribute to Daniel Gile*, Gyde Hansen, Andrew Chesterman, and Heidrun Gerzymisch-Arbogast (eds), Amsterdam, John Benjamins: 159–77.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli and Mariachiara Russo (2005) "Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus)", *Meta* 50, no. 4. **http://id.erudit.org/iderudit/019850ar**

Riccardi, Alessandra (2005) "On the Evolution of Interpreting Strategies in Simultaneous Interpreting", *Meta* 50, no. 2: 753–67.

Rosato, Enrica (2013) *Adjective Order in English: A Semantic Account with Cross-linguistic Applications*, Pittsburgh, Carnegie Mellon University.

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*, Francesco Straniero Sergio and Caterina Falbo (eds) Bern, Peter Lang: 53–90.

Sandrelli Annalisa, Claudio Bendazzoli and Mariachiara Russo (2010) “European Parliament Interpreting Corpus (EPIC): Methodological Issues and Preliminary Results on Lexical Patterns in SI”, *International Journal of Translation* 22, no. 1–2: 165–203.

Schjoldager, Anne (1995) “An Exploratory Study of Translational Norms in Simultaneous Interpreting: Methodological Reflections”, *Hermes, Journal of Linguistics* 14: 65–87.

Shlesinger, Miriam (2003) “Effects of Presentation Rate on Working Memory in Simultaneous Interpreting”, *The Interpreters’ Newsletter* 12: 37–49.

SSLMITDev **http://sslmitdev-online.sslmit.unibo.it/corpora/corporaproject.php?path=E.P.I.C**. (accessed 12 June 2016).

Sunnari, Marianna (1995) “Processing Strategies in Simultaneous Interpreting: “Saying it all” vs. Synthesis” in *Topics in Interpreting Research*, Jorma Tommola (ed.), Turku, University of Turku – Centre for Translation and Interpreting.

Toury, Gideon (1980) *In Search of a Theory of Translation*, Tel Aviv, Tel Aviv University, The Porter Institute for Poetics and Semiotics.

Van Dam, Ine Mary (1989) “Strategies of Simultaneous Interpretation” in *The Theoretical and Practical Aspects of Teaching Conference Interpretation: First International Symposium on Conference Interpreting at the University of Trieste*, Laura Gran and John Dodds (eds), Udine, Campanotto: 167–76.

Unpublished materials and grammar references

Barbafina, Silvia (2003) *Gestione delle stringhe aggettivali nell'interpretazione simultanea dall'inglese*, MA diss., University of Bologna.

Cappelli, Rita (2014) *L'interpretazione simultanea dal polacco all'italiano. Le strategie per affrontare le catene nominali*, PhD diss., University of Bologna.

Collins English Dictionary, definition of *more* **http://www.collinsdictionary.com/dictionary/english/more?showCookiePolicy=true** (accessed 21 April 2016).

Collins English Dictionary, definition of *most* **http://www.collinsdictionary.com/dictionary/english/most?showCookiePolicy=true** (accessed 21 April 2016).

Ghiselli, Serena (2015) *Le sfide traduttive dei sintagmi nominali con modificatori in posizione prenominale nell’interpretazione simultanea dall’inglese in italiano: uno studio sul corpus EPIC,* MA diss., University of Bologna.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*, Harlow, Longman.

Renzi, Lorenzo, Giampaolo Salvi, and Anna Cardinaletti (eds) (2001) *Grande grammatica italiana di consultazione*, Bologna, Il Mulino.

Swan, Michael (2005) *Practical English Usage*, Oxford: Oxford University Press.

"The translation challenges of premodified noun phrases in simultaneous interpreting from English into Italian", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.

Stable URL: http://www.intralinea.org/specials/article/2322

"On anaphoric pronouns in simultaneous interpreting", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2321

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy. This printout was generated directly from the online version of this article and can be freely distributed under Creative Commons License **CC BY-NC-ND 4.0**.

On anaphoric pronouns in simultaneous interpreting

By Ana Correia (University of Minho, Portugal)

Abstract & Keywords

English:

The successful establishment of anaphoric links between pronouns and their antecedents is a basic condition to ensure that a text is both cohesive and coherent. This is sometimes difficult to achieve when dealing with spoken texts and severe temporal restrictions as is the case in simultaneous interpreting. The present study focuses on personal and demonstrative pronouns. It is based on a random sample of transcripts of speeches and interpretations delivered at plenary sessions of the European Parliament (EP), taken from a larger pool of data, which will be included in an interpreting corpus to be compiled at the University of Minho.

Keywords: simultaneous interpreting, cohesion, anaphora, personal pronouns, demonstrative pronouns

Introduction

Research on anaphora has gained momentum in recent years due to the interest it has raised among scholars of Natural Language Processing and Artificial Intelligence, who became intent on finding solutions to resolve the ambiguities posed by this intriguing linguistic phenomenon. In the field of translation and interpreting (T&I) studies, anaphora is but a marginal topic of research. This may be because this phenomenon does not lend itself readily to observation, especially not in simultaneous interpreting (SI). Studying anaphora involves studying the relationship between two or more elements (i.e. the antecedent and the anaphor(s)), which are sometimes not easily identifiable in a text. This is all the more true if said text is translated simultaneously, in which case the likelihood increases that one of the elements in the anaphoric chain will be lost. Additionally, in simultaneous interpreting, the study of anaphora – and in fact of any other linguistic phenomenon – is further compounded by the need to transcribe the oral data beforehand. However, the developing paradigm of corpus-based interpreting studies may help to contravene this tendency, shedding light on such phenomena, which are crucial to achieve a deeper understanding of the mechanics behind discourse production. In the first section of this paper, we will briefly trace the evolution of corpus-based interpreting studies and provide an overview of some of the interpreting corpora that are currently available online. This section further describes the general outline of a teaching experience conducted at the University of Minho in connection with the compilation of an interpreting corpus. The second section deals with the object of study, i.e. anaphoric pronouns, and its relevance for simultaneous interpreting, which we will attempt to demonstrate in the third section by transcribing and analyzing some examples taken from a small sample of speeches. The fourth section provides some conclusions based on the findings discussed in the previous section.**[1]**

1. Corpus-based interpreting studies

Interpreting is a multi-faceted phenomenon which can – and no doubt must – be studied from a wide array of perspectives (Pöchhacker 2015). Indeed, since the 1950s, research on interpreting has been conducted under different paradigms. According to Moser-Mercer (1994), the fundamental distinction is that between the liberal arts paradigm and the natural science community. Most of the research conducted under the liberal arts paradigm was of a prescriptive and anecdotal nature, far removed from the realm of scientific experimentation. This paradigm corresponded to the first developmental stage of interpreting research. The first writings were essentially reports of personal experience and their main scope of application lied in training. Gradually, the scope broadened as did the methods employed to carry out research (Hale and Napier 2013). There was a growing concern with quantification and measurements, connected in particular with the surge of interest on the definition of quality in interpreting. This evolution was also accompanied by an important shift from prescriptive to descriptive research, which is now a cornerstone principle of the so-called liberal arts paradigm. For these reasons, it can be argued that the line between the liberal arts paradigm and the natural science community is becoming increasingly blurred. One of the factors that contributed to this state of affairs was without a doubt the advent of corpus linguistics, with its focus on the systematic and rigorous description of authentic data. Corpus linguistics was first applied to translation, yielding very successful results, and it became a popular research method among translation scholars ever since (Baker 1993, 1995; Laviosa 1998). In 1995, Susan Armstrong pioneered the idea of corpus-based interpreting studies (Armstrong 1995), which was further developed by Miriam Shlesinger (1998) in her seminal paper on the challenges and opportunities of extending corpus linguistics to interpreting studies. Following Shlesinger's call, many researchers did indeed venture into the compilation of interpreting corpora, some of which will be mentioned in the next subsection.

1.1. Interpreting corpora

In corpus linguistics, it is common to equate the notion of corpus with an electronic corpus available online through a dedicated search interface, which allows users to perform different types of queries. However, the word *corpus* can also be used to refer to any purposefully sampled body of texts, available either in paper or machine-readable form. This is indeed a broader conception of corpus, which is compatible with much of the work conducted in interpreting studies. In order to overcome problems of ecological validity, researchers in this field have been increasingly concerned with studying actual interpreting output and have therefore begun to build their own interpreting corpora. They are generally compiled by single researchers in the scope of their Master's or PhD projects, with many constraints. The compilation procedure begins with data collection, which could be a fairly simple retrieval of audiovisual files from a given online source or a more complex undertaking that requires the presence of the

researcher at a conference to record the proceedings and gather informed consents from all the participants. The data collected must then be transcribed according to a predefined set of transcription conventions. Researchers may rely on the help of speech-recognition software to produce draft versions of the transcripts but ultimately, whether it is done from scratch or not, transcription involves a great deal of manual labor, hence determining the limited size of many such corpora. Once data collection and transcription have been completed, the corpus is ready to be analyzed either manually or with the support of dedicated software. Not all of these ad hoc corpora result in electronic corpora. Nevertheless, both types of corpora are valuable sources of information for the study of interpreting and have contributed to achieve significant results in our field. For example, Setton (1999) based his cognitive-pragmatic approach on the analysis of such a corpus of English, German and Chinese transcripts. Naturally, these corpora are more useful if they are in machine-readable form, in which case they are fit to be processed using corpus linguistics tools. Nowadays there are countless free web-based tools that allow researchers to exploit their corpora (monolingual or bilingual) in meaningful ways, such as concordancers, taggers, aligners and terminology extractors.

It is widely acknowledged that the process of building interpreting corpora is a highly time-consuming task, mainly due to transcription work (Bendazzoli 2010b; Shlesinger 1998). However, over the past ten years, some scholars have successfully taken up the challenge of compiling such corpora. Among the first interpreting corpora was the European Parliament Interpreting Corpus (EPIC), built between 2004 and 2006 by an interdisciplinary team of researchers from the University of Bologna. EPIC is an on-line, trilingual corpus made up of speeches delivered at plenary sittings of the European Parliament in Italian, Spanish and English plus the respective interpretations (Bendazzoli and Sandrelli 2005; Russo et al. 2012). Italy has in fact been an active center for corpus-based interpreting research. In addition to EPIC, other corpora are worth mentioning, such as CorIT (media interpreting – consecutive and simultaneous) (Falbo 2012), DIRSI-C (conference interpreting - simultaneous) (Bendazzoli 2010a), and FOOTIE (media interpreting - simultaneous) (Sandrelli 2012). At the Hamburg Center for Language Corpora, affiliated with the University of Hamburg, several interpreting corpora have been compiled as well, representing not only simultaneous but also consecutive modes of interpreting (Bührig et al. 2012; House, Meyer and Schmidt 2012). One particular feature that distinguishes the work of researchers at Hamburg is that their repository includes community interpreting corpora, reflecting for example interpreter-mediated interaction in hospitals and in courtrooms. All around the world, the growing interest in corpus linguistics has spurred the creation of all sorts of different corpora suited to the study of a wide gamut of linguistic phenomena. Thanks in part to the pioneering work developed by the Language Resource Center for Portuguese, known as *Linguateca*, Portugal is no exception. We now have at our disposal a number of monolingual and multilingual corpora featuring Portuguese as either a source or target language such as *Corpus de Referência do Português Contemporâneo*. *Corpus do Português, CETEMPúblico*, *Le Monde Diplomatique*, *COMPARA*, *Per-Fide* and *OPUS*, to name but a few (for a comprehensive review of Portuguese corpora, see Berber Sardinha and Ferreira 2014). The last four examples are multilingual corpora with parallel alignment, hence particularly suited to research in translation studies. To our knowledge, the corpus mentioned in this study is an original attempt at building an interpreting corpus since, at present, there are no such corpora for (European) Portuguese. This is not surprising if we consider that spoken corpora in general are scarce and of limited size. However, members of the above-mentioned Hamburg Center for Language Corpora have created Dik - interpreting in hospitals corpus and CoSi - a corpus of consecutive and simultaneous interpreting, both of which include Brazilian Portuguese (Bührig at al. 2012; House, Meyer and Schmidt 2012). In Brazil, Luciana Ginezi is also compiling an interpreting learner corpus (Ginezi 2014). Due to the lack of interpreting corpora for European Portuguese, we decided to take up that challenge. We are currently involved in the compilation of the interPE corpus. It is a simultaneous interpreting multimedia corpus which includes Portuguese and English speeches delivered at the European Parliament plenary sittings. It will include not only the sentence-aligned transcripts of the original speeches and interpretations but also the corresponding audiovisual files. The interPE corpus will be composed of 20 Portuguese and 20 English speeches plus the respective interpretations, with each original speech averaging a duration of one and a half minutes. InterPE was envisaged as an open corpus, which means that more speeches will be added in the future. While it falls outside the scope of this paper to describe the compilation stages of our corpus, we would nevertheless like to report the involvement of some students of the University of Minho in the transcription stage, highlighting the potential benefits of this kind of work for the students.

1.2. Transcription: a teaching experience at the University of Minho

The study presented in this paper is part of a doctoral project, which in turn is based on a corpus of speeches delivered at EP plenary sittings by English and Portuguese MEPs (Members of the European Parliament) as well as the respective interpretations, in simultaneous mode. This initiative, which began in 2013-14, has been developed in collaboration with students attending the course unit of Principles of Interpreting, from the 3rd year of the undergraduate degree in Applied Languages of the University of Minho (Braga, Portugal). The students transcribed and/or revised speeches using *EXMARaLDA – Partitur*. The speeches were orthographically transcribed. The decision was made not to transcribe paralinguistic features such as pauses, hesitations, vowel lengthenings or false starts (among others), as this fell outside the scope of our study, which was exclusively concerned with anaphoric relations. The students further aligned the speeches with the respective interpretations using the web-based aligner *YouAlign*[2]. Each student was then asked to produce an analysis of the interpretations they transcribed, based on a typology adapted from Falbo (1998). We implemented a two-stage approach which allowed us to successfully involve students not only in the actual compilation of the corpus but also in the analysis of the data, as illustrated in Figure 1 below:

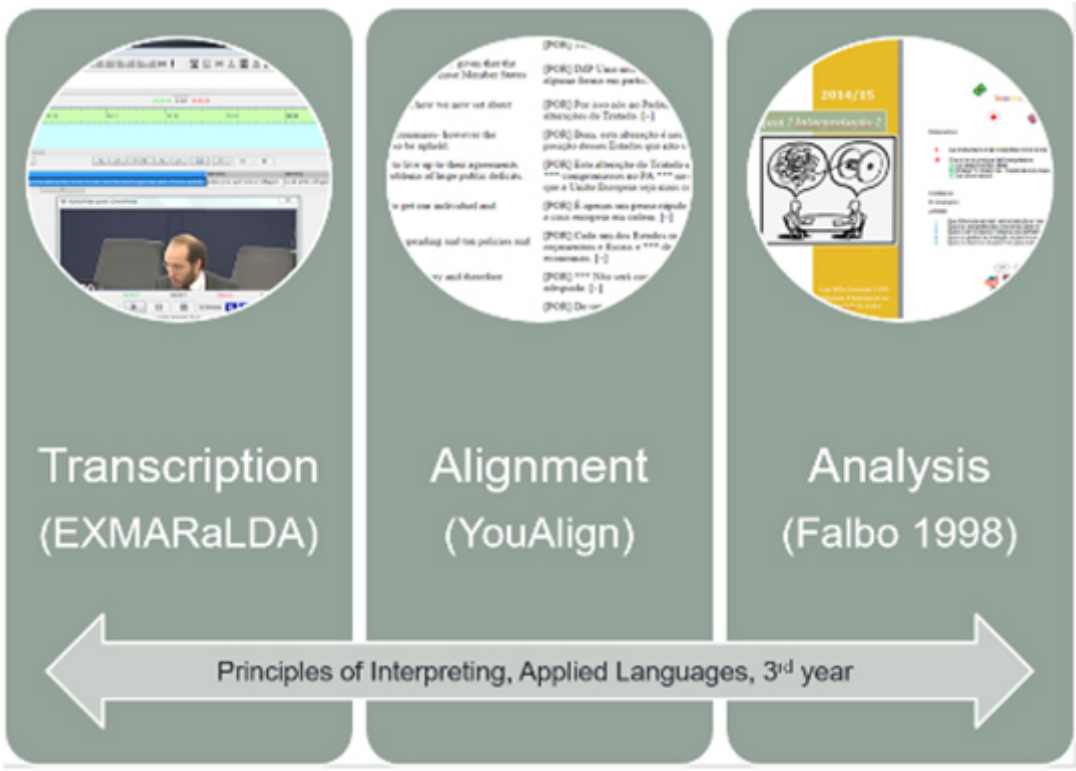


Figure 1: Methodology used in class.

This teaching experience led us to believe that such an approach produces positive results as it encourages students to acquire technical (connected with the early stages of corpus compilation) and analytical (connected with reasoning abilities and linguistic analysis) skills. According to the students' feedback, this exercise had a satisfactory outcome. In general, students claimed to have acquired a relevant and diverse set of skills that can actually help them in their future language-related careers. For example, among the benefits they mentioned were learning to use transcription and alignment software, and learning simple yet effective linguistic terminology to describe with scientific rigor some of the phenomena encountered in the speeches they transcribed. Incidentally, for the majority of students, who saw interpreting as a mission impossible, the analysis of authentic output contributed to demystify the work of the interpreter. It helped students gain a better understanding of simultaneous interpreting in the context of EP plenary sittings, drawing their attention to the delivery rates as well as to the syntactic and semantic complexity of the speeches.

2. Anaphoric pronouns in simultaneous interpreting

Anaphora is one of many linguistic devices employed by speakers to create "texture" or "textuality", that is, the property of "being a text". According to de Beaugrande and Dressler (1981), texts are communicative occurrences that must comply with seven principles of textuality: cohesion, coherence, intentionality, acceptability, informativity, situationality and intertextuality. Textuality is the property that allows a text to be acknowledged as such, rather than of a heap of disconnected sentences. The first two standards are text-centered. Cohesion, in particular, concerns grammatical dependencies. Coherence, in turn, depends upon the semantic connection between the sentences that make up a text or, in other words, it depends on whether they conform to our view of the world and on their adequacy to the communicative context. Anaphora is located at the level of cohesion. It is a lexico-grammatical mechanism that enables the establishment of referential chains, presupposing the existence of a referentially dependent element (the anaphor) which can only be interpreted in connection with another item that is present in the cotext. An anaphora can be coreferential if both elements designate the same real world entity, or non coreferential if they have different referents. There are also various types of anaphora, depending on the grammatical nature of the elements involved: pronominal, nominal, verbal, and adverbial (Lopes and Carapinha 2013; Charolles, 2002).

While anaphora has been studied in a wide range of disciplines, within different frameworks (Branco, McEnery and Mitkov 2005), we are particularly interested in anaphora as a discourse-level phenomenon, as conceived in text linguistics and discourse analysis, and the challenges it brings for simultaneous interpreters. In Translation and Interpreting Studies, many scholars have been concerned with the question of how translators are able to recreate cohesive and coherent texts in the target language (Baker 2011; Hatim and Mason 1997; Neubert and Shreve 1992, among others). This line of research was often connected with the quest for translation and interpreting universals (Blum-Kulka 1986). Anaphoric reference is generally addressed as a marginal topic subsumed under the broad umbrella of cohesion and coherence. This is for example the case of Shlesinger's (1995) paper on cohesive shifts in SI, where anaphora is but one of the various devices dealt with by the author, and Gallina's (1992) study on the cohesion of political speeches, which looks not only into reference but also ellipsis, conjunctions and lexical cohesion. Friedel Dubslaff's (1993) paper on anaphoric retrieval in simultaneous interpreting is one of the few examples where anaphoric reference is regarded as a research topic that is worthy of interest on its own. Snelling (1992) also addresses a number of syntactic as well as semantic issues that should be taken into consideration when interpreting from Portuguese. With regard to syntax, one of the problems on which he focuses in more detail is choice of subject. Based on his corpus, Snelling found that most sentences in Portuguese did not begin with the subject. In such cases, he recommends that interpreters working into English always begin their sentences with a subject, following the linearity of the subject-verb-object structure. As we will see in the next section of this paper, the search for a subject is a frequent obstacle faced by interpreters working from Portuguese into English, who often use pronouns to fill in that gap, often generating ambiguities and even erroneous chains of reference. These authors share an interest in cohesive ties and acknowledge their relevance for interpreting (see for example Gumul 2012). In

particular, anaphoric ties are a basic condition for the successful construction of any text, helping to ensure cohesion and coherence. Pronouns can be used to build anaphoric chains made up of not only intrasentential but also intersentential connections that can spread through an entire speech. Such a complex architecture may be costly in terms of processing requirements, and if anaphoric links are not properly established, that may well affect a text's communicative intelligibility. The cognitive processing underlying the mechanisms of reference building becomes more complex when it is conducted only in the spoken mode and under severe temporal restrictions as is the case in simultaneous interpreting. Thus, if we consider that a text results from the intersection of several anaphoric chains, it becomes clear that the study of anaphora is relevant for interpreting, which aims to ensure that a source text is rendered in the target language in a cohesive and coherent manner.

3. Empirical analysis of anaphoric pronouns in simultaneous interpreting

In this section, we present a small-scale exploratory study about anaphoric pronouns in simultaneous interpreting from Portuguese into English. We will begin by providing the frequency and distribution of the pronouns. This will be followed by a detailed analysis of examples taken from the corpus.

3.1. Frequency and distribution

For this empirical analysis, we selected a random sample of seven Portuguese speeches (plus English interpretations). The sample – to which we will refer as corpus throughout the remainder of this paper – was deliberately small in order to allow for a more in-depth qualitative analysis of the relevant examples. By randomly selecting the speeches, we ensured that the sample would be unbiased by any speech- (for example, topic), speaker- (for instance, gender) or interpreter-related variables (for example, professional experience) and that it would be free from researcher bias. The following exclusion criteria were applied before we could proceed with the random sampling to ensure that the data was homogenous in the first place:

> speakers who did not speak in their mother tongue (that is Portuguese);
>
> speeches by Commissioners and other non-MEP entities, which tend to be either long interventions that far outlast those of MEPs or very brief announcements of who has the floor;

We then proceeded to the extraction of all the pronouns, in the originals as well as in the interpretations. Since the corpus had been previously annotated with part-of-speech tags, the extraction process was completed semi-automatically, only requiring manual verification in a few ambiguous cases. The pronouns were then organized and counted according to type. The large majority of occurrences found in the corpus were of personal, relative and demonstrative pronouns, in that order. No possessive pronouns were found but we thought it relevant to extract all possessive determiners since these markers are often implicated in anaphoric relations, leading to ambiguous readings. Possessive determiners ranked third, after the personal and relative pronouns. The results are shown in table 1 below:

	PT (original)	EN (interpretation)	TOTAL
Personal pronouns	36	121	157
Relative pronouns	48	26	74
Possessive determiners	29	20	49
Demonstrative pronouns	13	22	35
TOTAL	126	189	315

Table 1: Number of pronouns per language and type.

This study focuses on personal and demonstrative pronouns. Personal pronouns were chosen because of the high number of total occurrences, which is more than twice the number of occurrences found for the second most frequent category of pronouns (namely, relative). Despite being the least frequent, demonstrative pronouns were chosen because of their potentially resumptive value, which allows for these pronouns to select wider antecedents. After extracting the pronouns, it was necessary to mark those that were part of anaphoric chains (that is the pronouns with anaphoric value). We found that: out of 36 personal pronouns, 26 were anaphoric; and out of 13 demonstrative pronouns, all 13 were anaphoric. This quantitative extraction was carried out only in the original speeches. We then copied into a spreadsheet all the occurrences of anaphoric pronouns in Portuguese in their extended context along with the respective interpretations in English. This allowed us to comparatively analyze the originals and the interpretations, focusing on how the anaphoric chains were rendered in the interpretations.

3.2. Personal and demonstrative pronouns

In the following sections, we present a non-exhaustive selection of cases where the anaphoric chains present in the original speeches were omitted or reformulated in the interpretations. While such operations could affect coherence, it is not our goal here to assess the extent to which the interpreted speeches are affected by the omission and/or reformulation of the anaphoric chains. The examples were taken from five out of the seven speeches analyzed; the originals are preceded by the acronym OS and interpretations by the acronym INT, each accompanied by the speech number (according to the chronological order in which they were delivered).

3.2.1 Personal pronouns

Corpus analysis uncovered a clear asymmetry between English and Portuguese in terms of the number of pronouns, as can be seen from Table 1. This is especially visible in the case of personal pronouns, which may be explained by the fact that English, as opposed to Portuguese, does not accept null subjects. This can further be explained by the tendency observed in the English interpretations toward paratactic structures. Example (1a) is representative of the speeches delivered by Portuguese MEPs, who tend to include several embedded clauses:

OS_2

(1a) Neste debate não podemos esquecer que existe uma proposta dum chamado pacto de competitividade
In this debate we cannot forget that there is a proposal of a so-called competitiveness pact

através do qual o diretório, comandado pela Alemanha, quer desferir novos ataques ao regime público
through which the directory, led by Germany, wants to launch new attacks against the solidary and

solidário e universal da segurança social, aumentar a idade da reforma e desvalorizar salários, tentando
universal public regime of social security, increase the age of retirement and devalue salaries, trying to

pôr fim à sua indexação à taxa de inflação apenas para beneficiar o setor financeiro, **o qual** pretende
put an end to their indexation to the inflation rate only to benefit the financial sector, **which** *intends to*

encontrar nas pensões novas formas de maiores ganhos especulativos.
find in pensions new ways to greater speculative gains.

INT_2

(1b) We must remember in this debate that there is a proposal relating to the competitiveness pact. Germany in particular seems to be very ready to attack the system of public security by lowering salaries, by exacerbating inflation largely for the benefit of the financial sector **and we know** that the financial sector wants to continue to gamble through private financing of pensions.

The complexity of the original speech takes its toll on the interpreter, who attempts to chunk the incoming message into smaller, more manageable bits of information. These chunking operations result in the creation of shorter sentences, to which the interpreter must assign a subject, as required by English grammar. According to Gile (1994: 48), however, 'a deviation from the source language structure may mean the interpreter is controlling the situation, whereas the selection of target language structures similar to source language structures indicates that the interpreter may be short of processing capacity'. As mentioned above, syntactic complexity is often the hallmark of Portuguese speeches. This factor is further compounded by the delivery rates, which in our corpus ranged between 140 and 181 words per minute (average = 156 wpm). These factors hinder the process of recognizing and assigning a syntactic subject to the new sentences. In such cases, our corpus analysis has shown that interpreters often resort to the generic personal pronoun "we". In (1b), this pronoun provides the interpreter with a plausible subject for the independent clause he**[3]** creates in his rendition. The use of "we" also proved a particularly useful instrument when the interpreter struggled to identify the antecedent in an anaphoric chain. In (2b) we have a rather unclear anaphoric link between the pronoun 'ela' and the immediately preceding antecedent ('Europa pós-queda do muro'):

OS_3

(2a) Isto parece-me uma perversão fundamental dos princípios da Europa pós-89, da Europa pós-queda do
This seems like a fundamental perversion of the principles of Europe after 89, of Europe after the fall of

muro. O que **ela** queria dizer é que nós não abandonaríamos os nossos irmãos europeus de qualquer país à
the Wall. What **she** *meant is that we would not abandon our European brothers of any country to*

censura e à repressão à liberdade de expressão.
censorship and to repression of freedom of expression.

INT_3

(2b) I think this runs completely counter to the principles of Europe, particularly after the fall of the Wall. **We** have said that we will not leave Europeans subject to censorship in any country.

Unable to identify the antecedent of 'ela', the interpreter has to look for an alternative that allows her to do away with the anaphoric chain without severely detracting from the coherence of her speech. In their efforts to segment the incoming speech into manageable chunks of information that allow them to stick to the canonical sentence order (subject-verb-object(s)), interpreters seem more prone to employ paratactic structures rather than hypotactic ones. In order to convert hypotaxis into parataxis, the interpreter is required to produce a syntactic subject. Since it is not always simple to come up with an appropriate subject, the use of the pronoun "we" acquires a strategic dimension of considerable usefulness. A great deal of attention has also been devoted to the study of "we" markers (we, us, our) in the specific context of European Parliament interpreting as a means of highlighting ideological assumptions (Beaton 2007; Dumara 2015). As we have seen in (2), the pronoun "we" can be used as an alternative to solve such anaphora-induced difficulties but there are other pronouns that can fulfill the same role, such as "they":

OS_6

(3a) O desaparecimento de Ai Weiwei tem de ser entendido no contexto do aumento desesperado da
The disappearance of Ai Weiwei has to be understood in the context of the desperate increase of

repressão política por parte das **autoridades Chinesas**. **Tudo** por medo de que o espírito revolucionário
political repression on the part of **Chinese authorities**. *All out of fear that the revolutionary spirit*

no mundo Árabe infete a **sociedade chinesa**.
in the Arab world might infect the **Chinese society**.

INT_6

(3b) The disappearance of Ai Weiwei has to be understood in the context of the tightening up of political repression **in China**. **They** are afraid that the democratic spring in the Arab world might infect **them**.

In (3b), the interpreter creates an anaphoric relation that did not exist in the original speech ('they'…'them'). In the first sentence, he replaces the agent ('por parte das autoridades chinesas') with a simple locative phrase ('in China'). This phrase becomes the antecedent for the pronoun at the beginning of the following sentence. Owing to the vagueness of this antecedent, which could refer to the Chinese authorities (as in the original), to the population or any other Chinese entity, the interpreter opts for 'they' to fill in the subject role, possibly due to the reminiscence of a plural antecedent uttered in the original speech ('autoridades chinesas'). This anaphoric chain has a third link ('them'), which necessarily follows from the interpreter's previous choice. It remains unclear though whether the two anaphors are coreferential. In any event, by using 'they', the interpreter leaves it up to his listeners to decide whether to interpret these two anaphors coreferentially and to determine what their referent(s) is(are).

3.2.2. Demonstrative pronouns

In addition to personal pronouns, demonstrative pronouns have also been found to serve as a strategic device when used resumptively. As already mentioned above, delivery rate is an essential variable in interpreting, which may lie at the origin of various comprehension and production problems encountered by interpreters, especially in simultaneous mode. In the specific context of the EP plenary sittings, which is notorious for rigid constraints on

floor allocation rules, delivery rates are often found to exceed the optimal threshold**[4]**. For that reason, most speakers prepare their speeches in advance, which means that they are generally closer to the literate pole of the oral-literate continuum (Shlesinger 1989). Adding to an already tense situation, speeches are typically encumbered by intricate lines of reasoning that are often hard to follow even for native speakers, as is the case in the excerpt transcribed in example (4a):

OS_2

(4a) Neste debate não podemos esquecer que existe uma proposta dum chamado pacto de competitividade
In this debate we cannot forget that there is a proposal of a so-called competitiveness pact

através do qual o diretório, comandado pela Alemanha, quer desferir novos ataques ao regime público
through which the directory, led by Germany, wants to launch new attacks against the solidary and

solidário e universal da segurança social, aumentar a idade da reforma e desvalorizar salários, tentando
universal public regime of social security, increase the age of retirement and devalue salaries, trying to

pôr fim à sua indexação à taxa de inflação apenas para beneficiar o setor financeiro, o qual pretende
put an end to their indexation to the inflation rate only to benefit the financial sector, which intends to

encontrar nas pensões novas formas de maiores ganhos especulativos. Queremos aqui manifestar a nossa
find in pensions new ways to greater speculative gains. We want here to manifest our

clara oposição a este caminho da integração europeia construído na base de políticas anti-sociais a que
clear opposition to this road of European integration built on the basis of antisocial policies to which

lamentavelmente este relatório dá cobertura ao apoiar o Livro Verde da Comissão Europeia, ao
this report regrettably gives credit by supporting the Green Book of the European Commission, by

admitir uma ligação da idade legal da reforma à esperança de vida e incentivar a permanência por um
allowing a connection between the age of retirement and life expectancy and encouraging permanence

período mais longo no mercado de trabalho, ao não excluir o apoio a sistemas de reformas privados
for a longer period of time in the job market, by not excluding support to private retirement systems

mesmo quando já se conhecem consequências graves da sua utilização especulativa por fundos e
even when there are already known consequences of their speculative use by private banks and funds

bancos privados que deixaram os idosos, designadamente mulheres idosas, na pobreza.
that have left the elderly, namely elderly women, in poverty.

INT_2

(4b) We must remember in this debate that there is a proposal relating to the competitiveness pact. Germany in particular seems to be very ready to attack the system of public security by lowering salaries, by exacerbating inflation largely for the benefit of the financial sector and we know that the financial sector wants to continue to gamble through private financing of pensions. I think we have to be clear about the dangers of **that**. These are antisocial policies and I think it's crucial that we understand **that**. We have to take account of increasing life expectancy and the fact that in many instances people are working for much longer periods. The document also talks about the intervention of the private sector but there are serious speculative risks related to that because many older women are being driven into poverty by this combination of circumstances.

It would seem that the interpreter was not able to keep up with the original speech. In particular, after the first sentence he was thrown off the track and forced to deploy 'coping tactics' (Gile 1995: 191). In this case, his tactics consisted in the use of the demonstrative pronoun "that" as a resumptive, which is taken to refer back to the preceding clauses. This allowed him to save time, leaving it to the listeners to put together the intended meaning of the speaker. Irrespective of all the compounding difficulties inherent to simultaneous interpreting, interpreters must see their renditions through, resorting to alternatives that do not always yield the best results. Although resumptive pronouns do offer a valid non-committal strategy, their intrinsic vagueness could impair the listeners' understanding of the interpreter's rendition. It can be argued, however, that the listeners, who are assumed to possess some degree of familiarity with the topics discussed at EP plenary sittings, may be able to fill in any semantic gaps on the basis of their previous knowledge. This speaks to the importance of extralinguistic factors in interpreting, namely, the listener's cognitive complements (Seleskovitch and Lederer 1984), which allow them to fill in gaps caused by any omissions or inaccuracies on the part of interpreters.

We have already mentioned above that, when interpreting from Portuguese into English, there is a recurrent use of parataxis to the detriment of hypotaxis, which forces interpreters to assign a syntactic subject to each new sentence or clause. When in doubt about an adequate subject, it was found that interpreters often used the pronoun "we" or "they". However, other evidence from the corpus showed that demonstrative pronouns can also be used for the same purpose, as in (5b):

OS_7

(5a) Obrigada, Senhora Presidente. Num tema necessariamente vasto, queria aqui deixar apenas dois
Thank you, Madam President. In a necessarily vast theme, I would like here to leave just two

breves apontamentos. O primeiro para chamar a atenção para os fatores de ameaça que hoje pesam sobre
brief notes. The first to draw attention to the factors of threat that today weigh over

inúmeros ecossistemas florestais.
numerous forest ecosystems.

INT_7

(5b) Thank you, Madam President. **This** is a very vast topic but I'd simply like to make two points. The first is that I'd like to draw your attention to the threats for forestry resources and then the exotic species that escape forest fires.

This excerpt was taken from the beginning of the speech. The original sentence was converted into two coordinate clauses joined by an adversative conjunction ('but'). Due to this segmentation, the interpreter had to find a subject for the first clause, which is the pronoun 'this'. This transformation makes more explicit the restrictive relationship

between the hypernym ('vast topic') and the hyponym ('two points'). In this case, the pronoun 'this' takes on a cataphoric value, unlike the pronoun 'that' in the following example which is both anaphoric and cataphoric:

OS_1

(6a) É condição sine qua non que a Líbia permita que o Alto Comissariado das Nações Unidas para os
It is a sine qua non condition that Libya allows the United Nations High Commissioner for

Refugiados volte a operar no país com um mandato alargado. Atrevo-me a dizer claramente: sem
Refugees to once again operate in the country with an extended mandate. I dare say clearly: without

ACNUR não há acordo.
UNHCR there is no agreement.

INT_1

(6b) The condition sine qua non is that Libya allows the UNHCR to come back to the country with an amplified agreement. I have to say **that** quite clearly: without UNHCR, no agreement.

In (6b) the pronoun 'that' resumes the idea conveyed in the previous sentence and, at the same time, it announces the reasoning that follows it. It would seem that, by placing the pronoun in a cataphoric position, the relationship between the anaphor and the postcedent becomes even more evident than in the original. In both (5b) and (6b), the interpretations have a higher degree of explicitness – the first due to the segmentation and the second due to the addition of the demonstrative.

4. Conclusions

Our corpus analysis has shown that English target speeches (the interpretations) globally have more pronouns than Portuguese source speeches (the originals). This is partly because interpreters are more prone to use hypotactic structures. In order to deal with this, interpreters tend to segment the input into small chunks and in doing so they are left with coordinated clauses to which they must assign suitable subjects. It was found that interpreters resort to pronouns such as 'we' and 'they' as a means of fulfilling that grammatical requirement. Our corpus analysis also showed that demonstrative pronouns were employed in anaphoric relations with a resumptive function, referring back to strings of embedded clauses as in (4b). Demonstrative pronouns further contributed to make more explicit the semantic logic that was only implicit in the original speeches, as was the case in examples (5) and (6). These findings suggest that personal and demonstrative pronouns are strategically used by interpreters to meet a grammatical requirement of the target language as a result of chunking operations. Additionally, these pronouns provide a non-committal alternative which is valuable to interpreters in case of doubt. However, the drawbacks of pronoun use can quickly overshadow the benefits if the intrinsic vagueness of pronouns prevents listeners from being able to identify the antecedent in the anaphoric relationship of which they form part. Although listeners are assumed to bring their previous knowledge to the context of simultaneous interpreting, that may not always be sufficient to overcome the vagueness introduced by some pronominal anaphors.

This type of study is based on the premise that reflection on the practice of interpreting through the analysis of authentic data, that is, a corpus (electronic or not), can promote the students' metalinguistic awareness, helping them to develop anticipation and problem-solving strategies (Sandrelli 2010). We have seen that there are a few corpora of interpreting but certainly not nearly as many as there are for (written) translation. However, it is fair to claim that corpus-based interpreting studies is gaining ground all around the world. Scholars engaged in this kind of research are well aware of the difficulties of creating interpreting corpora so, in line with the rationale behind the 1st Forlì International Workshop on Corpus-based Interpreting Studies, it is important that researchers join efforts in the future with regard to greater standardization of interpreting corpora, thus contributing to significant increases in terms of sheer size and representativeness. To the best of our knowledge, the interPE corpus is an original attempt at a simultaneous interpreting corpus featuring European Portuguese and, although it was created with the aim of studying anaphoric relations in simultaneous interpreting, we hope that it will come to serve a wider range of research purposes.

References

Armstrong, Susan (1995) "Corpus-based Methods for NLP and Translation Studies", *Interpreting* 2, no. 1/2: 141–62.

Baker, Mona (1993) "Corpus Linguistics and Translation Studies: Implications and Applications" in *Text and Technology: In Honour of John Sinclair*, Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds), Amsterdam/Philadelphia, John Benjamins: 233-52.

---- (1995) "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research", *Target* 7, no. 2: 223-43.

---- (2011) *In Other Words: A Coursebook on Translation*, London/New York, Routledge.

Beaton, Morven (2007) "Interpreted Ideologies in Institutional Discourse. The Case of the European Parliament", *The Translator* 13, no. 2: 271–96.

Bendazzoli, Claudio (2010a) *Il Corpus DIRSI: Creazione e sviluppo di un corpus elettronico per lo studio della direzionalità in interpretazione simultanea*, PhD diss., University of Bologna, Italy.

---- (2010b) *Corpora e interpretazione simultanea*, Bologna, Asterisco.

Bendazzoli, Claudio, and Annalisa Sandrelli (2005) "An Approach to Corpus-based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus)" in *MuTra – Challenges of Multidimensional Translation: Conference Proceedings*, Heidrun Gerzymisch-Arbogast and Sandra Nauert (eds), Saarbrücken, 1–12.

Berber Sardinha, Tony, and Telma São Bento Ferreira (eds) (2014) *Working with Portuguese Corpora*, London/New York, Bloomsbury Academic.

Blum-Kulka, Shoshana (1986) "Shifts of Cohesion and Coherence in Translation" in *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, Julianne House and Shoshana Blum-Kulka (eds), Tübingen, Narr: 17–35.

Branco, António, Tony McEnery, and Ruslan Mitkov (eds) (2005) *Anaphora Processing. Linguistic, Cognitive and Computational Modelling,* Amsterdam/Philadelphia, John Benjamins.

Bührig, Kristin, Ortrun Kliche, Bernd Meyer, and Birte Pawlack (2012) “The Corpus “Interpreting in Hospitals”: Possible Applications for Research and Communication Training” in *Multilingual Corpora and Multilingual Corpus Analysis*, Thomas Schmidt and Kai Wörner (eds), Amsterdam/Philadelphia, John Benjamins: 305–15.

Charolles, Michel (2002) *La référence et les expressions référentielles en français*, Paris, Ophrys.

de Beaugrande, Robert-Alain, and Wolfgang U. Dressler (1981) *Introduction to Text Linguistics*, London/New York, Longman.

Dubslaff, Friedel (1993) “Die Funktionen anaphorischer Proformen beim Simultandolmetschen aus dem Deutschen”, *Hermes, Journal of Linguistics* 11: 107–16.

Dumara, Barbara (2015) “How Can Interpreting Corpora Extend Our Knowledge on Intrusive ‘We’ in SI?”, Poster presented at the conference Corpus-based Interpreting Studies: The State of the Art. First Forlì International Workshop, 7-8 May 2015, University of Bologna at Forlì.

Falbo, Caterina (1998) “Analyse des Erreurs en Interprétation Simultanée”, *The Interpreters’ Newsletter* 8: 107–20.

---- (2012) “CorIT (Italian Television Interpreting Corpus): Classification Criteria” in *Breaking Ground in Corpus-based Interpreting Studies*, Francesco Straniero Sergio and Caterina Falbo (eds), Bern, Peter Lang: 157–85.

Gallina, Sandra (1992) “Cohesion and the Systemic-functional Approach to Text: Applications to Political Speeches and Significance for Simultaneous Interpretation”, *The Interpreters’ Newsletter* 4: 62–71.

Gerver, David (1969/2002) “The Effects of Source Language Presentation Rate on the Performance of Simultaneous Conference Interpreters” in *The Interpreting Studies Reader*, Franz Pöchhacker and Miriam Shlesinger (eds), London/New York, Psychology Press: 52–66.

Gile, Daniel (1994) “Methodological Aspects of Interpretation and Translation Research” in *Bridging the Gap. Empirical Research in Simultaneous Interpretation*, Sylvie Lambert and Barbara Moser-Mercer (eds), Amsterdam/Philadelphia, John Benjamins: 39–56.

---- (1995) *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam/Philadelphia, John Benjamins.

Ginezi, Luciana Latarini (2014) “Desafios para a Construção de um Corpus de Aprendizes de Interpretação Simultânea”, *TradTerm* 23: 165–91.

Gumul, Ewa (2012) “Variability of Cohesive Patterns. Personal Reference Markers in Simultaneous and Consecutive Interpreting.”, *Linguistica Silesiana* 33: 147-72.

Hale, Sandra, and Jemina Napier (2013) *Research Methods in Interpreting*, London/New York, Bloomsbury Academic.

Hatim, Basil, and Ian Mason (1997) *The Translator as Communicator*, London, Routledge.

House, Juliane, Bernd Meyer and Thomas Schmidt (2012) “CoSi - A Corpus of Consecutive and Simultaneous Interpreting” in *Multilingual Corpora and Multilingual Corpus Analysis*, Thomas Schmidt and Kai Wörner (eds), Amsterdam/Philadelphia, John Benjamins: 295–304.

Kleiber, Georges (1994) *Anaphores et pronoms*, Louvain-la-Neuve, Duculot.

Laviosa, Sara (1998) “The Corpus-based Approach: A New Paradigm in Translation Studies”, *Meta* 43, no. 4: 474-9.

Lopes, Ana C. M., and Conceição Carapinha (2013) *Texto, Coesão e Coerência*, Coimbra, Almedina.

Moser-Mercer, Barbara (1994) “Paradigms Gained or the Art of Productive Disagreement” in *Bridging the Gap. Empirical Research in Simultaneous Interpretation*, Sylvie Lambert and Barbara Moser-Mercer (eds), Amsterdam/Philadelphia, John Benjamins: 17–23.

Neubert, Albrecht, and Gregory M. Shreve (1992) *Translation as Text*, Kent, OH, The Kent State University Press.

Pöchhacker, Franz (2015) “Interpreting” in *Routledge Encyclopedia of Interpreting Studies*, Franz Pöchhacker (ed.), London/New York, Routledge: 197-200.

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*, Francesco Straniero Sergio and Caterina Falbo (eds), Bern, Peter Lang: 35-90.

Sandrelli, Annalisa (2010) “Corpus-Based Interpreting Studies and Interpreter Training: A Modest Proposal” in *Translationswissenschaft: Stand und Perspektiven. Innsbrucker Ringvorlesungen zur Translationswissenschaft* VI, Lew Zybatow (ed.), Peter Lang: 69–90.

---- (2012) “Interpreting Football Press Conferences: The FOOTIE Corpus” in *Interpreting across Genres: Multiple Research Perspectives*, Cynthia J. K. Bidoli (ed.), Trieste, Edizioni Università di Trieste: 78–101.

Seleskovitch, Danica, and Marianne Lederer (1984) *Intérpreter pour Traduire*, Paris, Didier Érudition.

Setton, Robin (1999) *Simultaneous Interpretation: A Cognitive-pragmatic Analysis*, Amsterdam/Philadelphia, John Benjamins.

Shlesinger, Miriam (1989) *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum*, M.A. diss., Tel Aviv University, Israel.

---- (1995) “Shifts in Cohesion in Simultaneous Interpreting”, *The Translator* 1, no. 2: 193–214.

---- (1998) “Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies”, *Meta* 43, no. 4: 486–93.

Snelling, David (1992) *Strategies for Simultaneous Interpreting from Romance Languages into English*, Udine, Campanotto.

Notes

[1] This study is part of a doctoral project, supported by grant no. SFRH/BD/88142/2012 and awarded by the Portuguese Foundation for Science and Technology under the Human Potential Operational Program. It is cofunded by the European Social Fund and the Portuguese Ministry of Education and Science.

[2] http://www.youalign.com/

[3] Thanks to the audio of the interpretations we were able to distinguish male from female interpreters, hence in this paper we use gender-marked pronouns to refer to the interpreters.

[4] According to Gerver (1969/2002) that would be in the range of 95 to 120 words per minute.

"On anaphoric pronouns in simultaneous interpreting", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2321

"Interpreting Universals: A study of explicitness in the intermodal EPTIC corpus", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2320

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy.

Interpreting Universals: A study of explicitness in the intermodal EPTIC corpus

By Niccolò Morselli (University of Bologna, Italy)

Abstract & Keywords

English:

This paper presents a study on explicitness in the European Parliament Translation and Interpreting Corpus (EPTIC). EPTIC (Bernardini et al. 2013) is a bilingual, bidirectional and intermodal corpus of EP plenary session speeches in English and Italian. It contains transcripts of both source speeches and their interpreted versions, as well as their written counterparts in the form of minutes and their translations. The study set out to test the findings of the quantitative analysis on explicitness in English interpretations carried out by Kajzer-Wietrzny (2012). The indicators of explicitness for the investigation of English (linking adverbials, apposition markers and optional that) were matched by comparable indicators for the investigation of Italian and applied to the relevant sub-corpora of EPTIC. First, a quantitative analysis was carried out, both from a monolingual comparable perspective (comparing speeches and interpretations in the same language), and from an intermodal perspective (comparing interpretations and translations). Second, a parallel qualitative analysis was performed. Some interesting differences according to language direction emerged, such as the Italian interpreters' preference to add apposition markers, or the tendency of English interpreters to leave out linking adverbials.

Keywords: explicitation, intermodal corpus, apposition markers, linking adverbials, simultaneous interpreting, interpreting universals, italian

1. Introduction

This study focuses on the thorny issue of translation and interpreting universals in general, and on the highly debated universal of explicitation in particular. Many translation scholars have so far questioned the existence of these concepts (Mauranen and Kujamäki 2004).

In translation studies, the so-called Explicitation Hypothesis was first put forward by Blum-Kulka (1986), who analysed explicitation from the perspective of discourse analysis, focusing on shifts in cohesion and coherence, and which led to subsequent studies. Much work has been done so far with the help of corpora to put this hypothesis to the test from many other perspectives.

After also Baker (1993 and 1996) listed it as a potential universal, explicitation has in fact been studied both as an S-Universal[1] in parallel corpora, for example, by Øverås (1998), and as a T-Universal, that is to say from a comparable monolingual perspective,for example by Olohan and Baker (2000) and Puurtinen (2004). These two perspectives have also been combined, for instance, by Pápai (2004) and Konšalová (2007). As for the different forms of explicitation, research has moved from Blum-Kulka's text cohesive elements to a wide variety of linguistic phenomena, as in Pápai and Konšalová's studies, or to the analysis of a single form of explicitation such as the optional *that* in Olohan and Baker's research.

Of the few studies on explicitation in interpreting studies, most are still based on very small collections of texts, mainly produced by advanced students of conference interpreting. This is the case of the study carried out by Schjoldager (1995/2002), the first study also focusing on explicitation in interpreted texts in the simultaneous mode, using parallel and intermodal corpora. From this first study on explicitation in simultaneous interpreting, no cases of the interpreter making explicit something implicit in the source speech emerged, while subsequent studies have on the whole confirmed Blum-Kulka's initial Explicitation Hypothesis. Shlesinger (1995) noticed that when the source text had an elliptical structure omitting an element mentioned before, the advanced interpreting students in her sample tended to repeat the missing element or to find a synonym, 'thus making the connection more explicit' (Shlesinger 1995: 201). Gumul (2006) maintains that an increased level of explicitness in the target text is often due to interpreting-specific factors, such as the interpreters' need to rephrase the utterance to add a new piece of information or to correct themselves. Gumul also noticed that sometimes her student interpreters tended to add words without adding information, filling a pause while waiting for the next piece of information to come. For this reason, she finds it difficult to maintain that these are examples of explicitation implied by the interpreting process itself. Ishikawa's study (1999) represents a rare example of investigations on explicitation in simultaneous interpreting carried out on texts interpreted by professional conference interpreters – in this specific case Japanese professionals working into English, that is not into their mother tongue, adding yet another variable to an already complex task. In her study, she singled out some cases of 'pure explicitations' thus confirming Blum-Kulka's hypothesis, but she also argued that in many cases the interpreters preferred implicitation (1999: 252).

As for universals in general, Chesterman pointed out that 'some have been corroborated more than others, and some tests have produced contrary evidence, so in most cases the jury is still out' (2004:39) and explicitation makes no exception. Other authors have also put Blum-Kulka's Explicitation Hypothesis to the test (Pym 2005; Becher 2010a, 2010b), focusing on explicitation in translation, and they reached the conclusion that explicitation as a universal should be regarded as a myth to be debunked because of the vagueness of the very concept of explicitation. Notably Pym (2005) explained this phenomenon within his model of risk aversion, and Becher (2010a and 2010b) maintained that the studies confirming the explicitation hypothesis featured serious weaknesses as for the methodology adopted, because they did not stick to the definition of explicitation as a phenomenon inherent to the translation process, but

included also different types of explicitation (as in Øverås 1998; Pápai 2004). Further criticism was raised by Becher (2010a) because the corpora used were unbalanced or because one could not have access to source texts (such as in Olohan and Baker 2000). He therefore suggested (2010b) that the assumption of explicitation being a universal should be discarded for good. Also Baumgarten, Meyer and Özçetin (2008) critically investigated explicitation both in translated and interpreted texts, and concluded that the increased level of explicitation in interpreted renditions from Portuguese into German of the term *Amazônia* were to be ascribed to other factors, such as interpreter's style, interpreting mode and other social and cultural variables (2008: 198).

To complete this short overview of this field of research, Kajzer-Wietrzny's corpus-based study (2012) on interpreting universals is important to mention for two main reasons. The first one is that in this study professional conference interpreters' performances are analysed. Secondly, this study was conducted using a larger corpus, namely the Translation and Interpreting Corpus (TIC) (2012: 57), which is one of the few intermodal corpora used in interpreting research so far. More specifically, it is an English monolingual comparable corpus including ten subcorpora in total. The oral part consists of one subcorpus of speeches in English and four subcorpora of interpreted versions into English of speeches given at the European Parliament from four different languages, namely Spanish, French, German and Dutch. The written part contains transcripts of the English speeches and of the interpreted versions of the speeches pronounced in the other languages, as well as further four subcorpora of texts translated into English from the same four other languages. In particular, in her study on interpreting universals, Kajzer-Wietrzny investigated explicitness, analysing linking adverbials, apposition markers and optional *that* as explicitness indicators.

Besides the weaknesses of the studies on explicitation in interpreting already referred to above – analysing principally small-scale collections of texts often produced by student interpreters – existing studies on explicitation in interpreting have applied extremely varied methodologies, thus making the findings considerably difficult to compare with other investigations of the same linguistic phenomenon.

Additionally, scholars have investigated explicitation not always clarifying the concept itself. As it emerges from this review, many have looked at explicitation (sometimes implicitly) defining it as the process of rendering covert information in the source text in an explicit way in the translated text. For instance, Baker (1996: 180) rather generally defined explicitation as 'the tendency to spell things out'. On the other hand, Kajzer-Wietrzny analysed explicitness, unequivocally defining it as a feature of target texts, rather than a strategy or a technique. In our study the term 'explicitness' is also adopted, since the subject of this investigation is not the process of making overt in the target text implicit information in the source text, but the tendency of target information of being more explicitly encoded, namely a textual feature not necessarily deriving from an explicitation process.

All these controversial aspects might also be some of the causes why clear and sound evidence of explicitation being an interpreting universal has not been yet been forthcoming. Nevertheless, explicitation is still considered one of the most popular 'candidates' in this quest. The investigation described in the following study is an initial attempt at searching for explicitness in Italian and English texts, both translated and interpreted (in the simultaneous mode). One of the aims of this study is to produce comparable results with future research projects, by adopting a quantitative approach, which is typical of corpus-based studies and has already been adopted in Kajzer-Wietrzny's study mentioned above. For this reason, in our study we applied the methodology used by Kajzer-Wietrzny (2012) in order to obtain comparable data, though working also on different language pairs.

In addition, a qualitative analysis was also performed to combine the invaluable contribution provided by corpora with a qualitative approach aimed at corroborating quantitative data, enriching it with further insights.

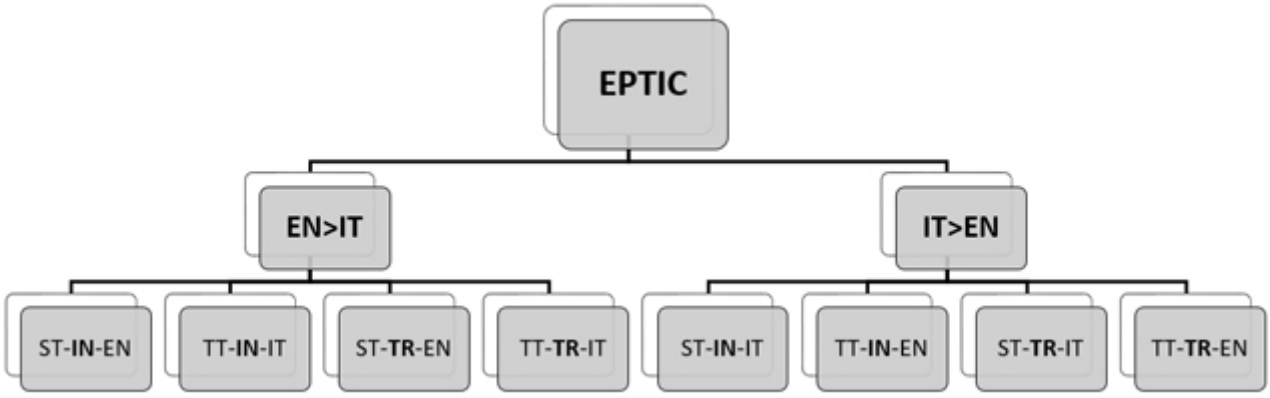
2. A study of explicitness in oral and written texts in English and Italian

2.1 Materials

The texts used for this study come from the *European Parliament Translation and Interpreting Corpus* (Bernardini, Ferraresi and Miličević 2013), a machine-readable multifaceted resource recently developed at the Forlì campus of the University of Bologna. EPTIC evolved from EPIC, the trilingual *European Parliament Interpreting Corpus* (Sandrelli and Bendazzoli 2005; Bendazzoli 2010) including speeches in Italian, English and Spanish and their interpretations in the same languages. Overall, EPTIC is an extension of EPIC with a difference - it does not contain Spanish subcorpora but it includes four subcorpora of written texts, both in Italian and in English. As a result, EPTIC is to date a unique bilingual, bidirectional and intermodal corpus of European Parliament plenary session speeches in English and Italian, containing both transcripts of source speeches and their interpreted versions paired with their written counterparts in the form of source-language minutes and their translations.

Plain text transcripts and headers containing extra-linguistic information (metadata) were taken from EPIC, while their respective minutes and their independently produced translations were downloaded from the official website of the European Parliament. Transcripts were POS-tagged and lemmatized**[2]**, and subsequently sentence-level alignment for parallel and intermodal pairs was carried out. For this study the corpus has been accessed through a UNIX server with a SSH client, and investigations were performed by using CQP syntax with a command line interface.

Thanks to its structure (Figure 1), EPTIC offers the key advantage of supporting investigations from various perspectives: its content can be examined from a comparable monolingual point of view, by comparing interpreted and translated subcorpora with oral and written source production in the same language, and from a monolingual intermodal perspective, by contrasting interpreted and translated texts, as well as from a bilingual parallel perspective that enables the investigation of target texts and their source texts.



ST: source texts
TT: target texts
IN: interpreted
TR: translated

IT: Italian
EN: English

ST-IN: transcripts of the original speech
TT-IN: transcript of the interpreter's rendition
ST-TR: minutes in the original language
TT-TR: translated minutes

FIGURE 1: Structure of EPTIC

As far as the corpus size is concerned, EPTIC subcorpora were not very much balanced in its original version (Table 1). For the purposes of this study, the first EPTIC enlargement was carried out adding transcripts, thus raising the number of tokens from 175,122 to 253,818 (Table 2).

SUBCORPUS	NUMBER OF TEXTS	TOTAL WORD COUNT	% OF EPTIC
ST-IN-EN	81	41,869	23.91
ST-TR-EN	81	36,685	20.95
TT-IN-IT	81	33,675	19.23
TT-TR-IT	81	36,876	21.06
ST-IN-IT	17	6387	3.65
ST-TR-IT	17	6234	3.56
TT-IN-EN	17	6577	3.76
TT-TR-EN	17	6819	3.89
TOTAL	**392**	**175,122**	**100**

TABLE 1: Size and composition of EPTIC (Bernardini, Ferraresi and Miličević 2013)

SUBCORPUS	NUMBER OF TEXTS	TOTAL WORD COUNT	% OF EPTIC
ST-IN-EN	81	41,869	16.50
ST-TR-EN	81	36,685	14.45
TT-IN-IT	81	33,675	13.27
TT-TR-IT	81	36,876	14.53
ST-IN-IT	61	26,088	10.28
ST-TR-IT	61	25,244	9.95
TT-IN-EN	61	26,113	10.29
TT-TR-EN	61	27,268	10.73
TOTAL	**568**	**253,818**	**100**

TABLE 2: Size and Composition of EPTIC after the first enlargement

2.2 Objectives and methods

Our study set out to test the findings of the quantitative analysis on explicitness in English interpretations and translations carried out by Kajzer-Wietrzny (2012). A second objective was to perform the first investigation into explicitness in Italian applying the same research design to the EPTIC Italian subcorpora. Furthermore, the study was also aimed at identifying tendencies in terms of higher or lower explicitness of interpreted and translated texts, in order to see whether some generalizations could be drawn.

Kajzer-Wietrzny's three indicators of explicitness for the investigation of English were borrowed for our study; they include the use of linking adverbials, apposition markers and optional *that,* since they are widely considered to be linguistic signals for an increased level of explicitness (Kajzer-Wietrzny 2012: 76). The set of linking adverbials comprised: *as a consequence, as a result, consequently, hence, in consequence, therefore* and *thus*. The apposition markers adopted indicated reformulations with an explicitation function: *that is, that is to say, to be (more) precise, to be (more) specific, to be exact, namely* and *in other words*. The third indicator, the optional *that* connective after reporting verbs, differed from the previous two because it indicates increased syntactic explicitness rather than

content reformulations. The reporting verbs chosen were: *admit, believe, claim, hope, know, suggest, say* and *tell.* All these indicators are exactly the same as in the study by Kajzer-Wietrzny.

For the investigation of Italian, the aforementioned indicators were matched by comparable Italian ones, identifying similar sets of linguistic elements characteristic of this language. So, for the purposes of this study, some adaptations have been made to the set of Italian indicators. In particular, an authoritative Italian grammar text was taken as a reference, specifically the *Grammatica Italiana* by Serianni, which provides lists of both linking adverbials and apposition markers (1989: 541-542). From these lists, archaic items such as *onde* and *laonde* were left out, since they are virtually never used today, neither in prestige high-register, while other elements, such as *per cui*, *di conseguenza* and *in altre parole*, were added in order to have the same number of elements in each set (English and Italian). The linking adverbial *infatti* was excluded, given its numerous pragmatic meanings and functions and its high frequency in spoken Italian language which might have distorted the results. The Italian linking adverbials selected were: *dunque, quindi, perciò, pertanto, sicché, per cui* and *di conseguenza*, while the set of apposition markers included: *cioè, ossia, ovvero, se vogliamo, vale a dire, per essere precisi* and *in altre parole.* As for the English optional *that*, this syntactic indicator was replaced with a similar one, as in Italian the connective after reporting verbs has always to be stated. In Italian, in fact, when the subordinate clause after reporting verbs has the same subject of the main clause, the speaker can choose between an explicit clause structure, using the connective *che* followed by the subject and the conjugated verb, and an implicit clause structure followed by the connective *di* and an infinitive verb, without having to repeat the subject. The reporting verbs chosen were: *affermare, aggiungere, ammettere, annunciare, asserire, avvertire, comunicare, confessare, considerare, credere, dichiarare, dire, dubitare, esclamare, informare, negare, promettere, raccontare, ricordare, rispondere, ritenere* e *trovare* (Serianni 1989: 552). In this subset of indicators, 22 elements have been analysed, as opposed to the only 8 items in English. In this case, we could not keep the same number of reporting verbs in English, since these reporting verbs are all currently in use in Italian.

Firstly, a quantitative macroanalysis**[3]** was carried out both from a monolingual comparable perspective, comparing speeches and interpretations in the same language, and from an intermodal perspective, comparing the occurrences of the selected sets of items in interpretations and translations. Secondly, a parallel qualitative microanalysis was performed to check if the trends of increased explicitness highlighted by the quantitative analyses were actually confirmed. The underlying tenet of the chosen methodology is that combining both a quantitative and qualitative approach gives researchers the opportunity to verify that the observations that emerged from the quantitative investigation are in fact the phenomenon really looked for, thus allowing sounder conclusions to be drawn.

2.3 Statistical significance test and limitations of the applied methodology

In order to verify the null hypothesis, that is that there were no differences between the raw frequencies of the investigated explicitness indicators, Fisher's exact test was applied and carried out thanks to *R*, a freely accessible software for statistical computing**[4]**. This test is expected to produce more precise *p*-values for not too large counts (Baayen 2008:113), as this is the case of the chosen items' frequencies, given the size of EPTIC and the kind of investigated linguistic phenomena**[5]**.

The first limitation of this methodology is related to the statistical significance test. To carry out Fisher's exact test, the number of the occurrences of the indicator (x) and the total number of tokens (N) of the investigated subcorpus must be inserted in the matrix (N – x). Since some indicators consist of more than one token, it is impossible to take this into account while performing the subtraction, otherwise each token of the same occurrence would count as separate occurrences. Therefore, every indicator was considered as if it were composed by a single token. It was thought that this does not affect the statistical relevance result.

Secondly, for an analysis to be as similar as possible in two different languages, first a preliminary contrastive analysis should be performed in order to verify that the sets of selected indicators are in fact exhaustive and equally represented in that language. Nevertheless, this would have led to a different methodology and to results not comparable with Kajzer-Wietrzny's study (2012).

Thirdly, it must also be acknowledged that it would have been better to use a corpus with perfectly balanced subcorpora, while some EPTIC components are larger than others, in spite of the efforts made within the scope of this study, as mentioned in 2.1.

3. Results

3.1 Introduction

In the following sections the results of the investigations conducted in the eight EPTIC subcorpora are grouped according to the three different sets of selected explicitness indicators, and presented from a traditional monolingual perspective. All the results are shown together with raw and normalized frequencies**[6]** and with the outcomes of Fisher's exact test for statistical relevance. First, the tendencies which emerged are discussed carrying out a comparable and an intermodal macroanalysis. For the purposes of the intermodal analysis, source subcorpora have also been statistically compared, to ensure that no significant differences due to external factors could affect the comparison between the target subcorpora, in terms of occurrences of the selected items, between original speeches and written minutes. Second, a parallel qualitative analysis is performed to confirm or disconfirm two quantitative tendencies that have emerged from the previous investigations and were selected as particularly interesting.

Even though a quantitative macroanalysis follows, the corpus size allowed us to ascertain that every single occurrence is precisely the linguistic phenomenon looked for.

3.2. Macroanalysis on English subcorpora

3.2.1 Frequency of explicitness indicators

Table 3 shows the results obtained from the different queries launched in the English oral and written subcorpora, both from a comparable and an intermodal perspective.

As regards **linking adverbials**, the data related to the oral subcorpora, namely original speeches and interpreted production into English, showed 57 occurrences in the interpreted subcorpus, normalized to 100,000 words, which was exactly the same number of normalized occurrences in comparable original speeches in English. As for written subcorpora, source texts (minutes) featured 65 occurrences of the seven selected items, while in the translated subcorpus the number of occurrences was almost three times as much, that is to say 183 in total. In this case, there were no statistically significant differences between the two subcorpora of oral and written source texts that could distort this comparison, counting respectively 57 and 65 occurrences.

As for **apposition markers**, the corresponding occurrences were 46 in interpreted texts and 26 in the comparable oral subcorpus. In written subcorpora, 51 apposition markers were counted in translated texts and 26 in comparable written texts, and therefore the normalized number of occurrences in the subcorpora of source texts is the same.

The results concerning the **optional** *that* connective after reporting verbs indicated that in every subcorpus this connective was verbalized in more than 50 per cent of cases. More precisely, the structure with optional *that* occurred in 78 per cent of the cases in the subcorpus of interpreted texts and in 70 per cent of cases in that of original speeches, while it appeared in 84 per cent of the cases in written translations and in 80 per cent of the cases in written texts.

From an intermodal perspective, it can be noted that the normalized occurrences of linking adverbials in the subcorpus of interpretations were 57, while they accounted for 183 in the subcorpus of translations. The apposition markers in the interpreted English subcorpus were 46 compared to the 51 occurrences in the translated English one. The data on optional *that* show that interpreters opted for the verbalisation of this connective in 78 per cent of cases, while translators did that in 84 per cent of cases.

LINKING ADVERBIALS ENGLISH				
SUBCORPUS	**st-in-en**	**tt-in-en**	**st-tr-en**	**tt-tr-en**
RAW FREQUENCY	24	15	24	50
NORMALIZED FREQUENCY (100,000 words)	57	57	65	183
STATISTICAL COMPARISON significance indicated by: **(*)**	**monolingual comparable**		**monolingual intermodal**	
	st-in-en vs. tt-in-en	**st-tr-en vs. tt-tr-en**	**st-in-en vs. st-tr-en**	**tt-in-en vs. tt-tr-en**
***p*-Value** significant with $p < 0,05$	1	**1.72×10^{-5} (*)**	0.6666	**2.852×10^{-5} (*)**
APPOSITION MARKERS ENGLISH				
SUBCORPUS	**st-in-en**	**tt-in-en**	**st-tr-en**	**tt-tr-en**
RAW FREQUENCY	11	12	9	1
NORMALIZED FREQUENCY (100,000 words)	26	46	26	51
STATISTICAL COMPARISON significance indicated by: **(*)**	**monolingual comparable**		**monolingual intermodal**	
	st-in-en vs. tt-in-en	**st-tr-en vs. tt-tr-en**	**st-in-en vs. st-tr-en**	**tt-in-en vs. tt-tr-en**
***p*-Value** significant with $p < 0,05$	0.2001	0.09197	1	0.8459
OPTIONAL *THAT*				
SUBCORPUS	**st-in-en**	**tt-in-en**	**st-tr-en**	**tt-tr-en**
RAW FREQUENCY	97/138	63/81	90/112	75/89
FREQUENCY IN PERCENT	70%	78%	80%	84%
STATISTICAL COMPARISON significance indicated by: **(*)**	**monolingual comparable**		**monolingual intermodal**	
	st-in-en vs. tt-in-en	**st-tr-en vs. tt-tr-en**	**st-in-en vs. st-tr-en**	**tt-in-en vs. tt-tr-en**
***p*-Value** significant with $p < 0,05$	0.2704	0.5792	0.3282	0.0793

TABLE 3: Results - English[7]

ST: source texts
TT: target texts
IN: interpreted
TR: translated

IT: Italian
EN: English

ST-IN: transcripts of the original speech
TT-IN: transcript of the interpreter's rendition
ST-TR: minutes in the original language
TT-TR: translated minutes

3.2.2 Major findings

The investigations carried out in the English subcorpora reveal an overall tendency towards an increased frequency of explicitness indicators in interpreted and translated subcorpora. The only exceptions are linking adverbials, whose frequency remains constant in oral subcorpora but increases in the translated subcorpus compared to the comparable written one. Only this very last tendency of higher explicitness marked by a larger number of linking adverbials in translated English texts is statistically significant according to Fisher's exact test. Table 4 summarises the explicitness tendencies that have emerged from the comparable monolingual analysis in English.

SUBCORPUS	**NORMALIZED**	**EXPLICITNESS**	**STATISTICAL**

		FREQUENCY	IN TARGET SUBCORPUS	SIGNIFICANCE
LINKING ADVERBIALS ENGLISH	tt-in-en	57	CONSTANT	NO
	st-in-en	57		
	tt-tr-en	183	**HIGHER (*)**	**YES**
	st-tr-en	65		
APPOSITION MARKERS ENGLISH	tt-in-en	46	HIGHER	NO
	st-in-en	26		
	tt-tr-en	51	HIGHER	NO
	st-tr-en	26		
OPTIONAL *THAT*	tt-in-en	78%	HIGHER	NO
	st-in-en	70%		
	tt-tr-en	84%	HIGHER	NO
	st-tr-en	80%		

TABLE 4: Explicitness tendencies – comparable perspective (English)

The English intermodal analysis has brought to light a clear tendency towards a higher degree of explicitness in translated production compared to interpreted texts, even though this tendency is statistically confirmed only in the case of linking adverbials. In no comparison between written and oral source subcorpora can a statistically significant difference be observed, thus confirming that potential tendencies are to be ascribed to the translating or interpreting process and not to differences already present in the source subcorpora. Table 5 indicates the frequencies of the three selected explicitness indicators in English from an intermodal point of view.

	SUBCORPUS	NORMALIZED FREQUENCY	EXPLICITNESS	STATISTICALLY SIGNIFICANT
LINKING ADVERBIALS ENGLISH	tt-in-en	57	**HIGHER (*) in tt-tr-en**	**YES**
	tt-tr-en	183		
	st-in-en	57	NO statistically significant difference between source subcorpora	
	st-tr-en	65		
APPOSITION MARKERS ENGLISH	tt-in-en	46	HIGHER in tt-tr-en	NO
	tt-tr-en	51		
	st-in-en	26	NO statistically significant difference between source subcorpora	
	st-tr-en	26		
OPTIONAL *THAT*	tt-in-en	78%	HIGHER in tt-tr-en	NO
	tt-tr-en	84%		
	st-in-en	70%	NO statistically significant difference between source subcorpora	
	st-tr-en	80%		

TABLE 5: Explicitness tendencies – intermodal perspective (English)

3.3 Macroanalysis on Italian subcorpora

3.3.1 Frequency of explicitness indicators

As was mentioned before (section 2.2), the three Italian explicitness indicators were chosen so as to be as similar as possible to the English ones. Since the Italian language does not have an equivalent connective to the optional *that*, but only an alternative construction with the connective preposition followed by the implicit form, this English syntactic explicitness indicator was replaced with the most similar Italian explicit structure. Table 6 summarises the results of the queries launched in the Italian subcorpora together with the respective statistical significance tests.

In the subcorpus of interpreted texts into Italian, 273 normalized occurrences of **linking adverbials** were counted, while there were 207 occurrences in the original Italian oral subcorpus. As for written subcorpora, in the translated one there were 106 occurrences of the seven selected items and in the comparable written subcorpus the number of occurrences is 136. Also in this case there were no statistically significant differences between the two subcorpora of oral and written source texts.

The investigations carried out for the second indicator, the set of **apposition markers**, showed that 51 normalized occurrences of this linguistic phenomenon could be identified in interpreted texts, but this frequency went up to 104 in the comparable oral Italian subcorpus. The same trend was observed in the two corresponding written subcorpora, where occurrences accounted for 30 in translated texts and 76 in comparable written ones. From the control comparison, in this case, no statistically significant differences were observed between source subcorpora either.

The frequencies obtained show that the third indicator chosen, namely the **explicit structure** after reporting verbs in clauses with the same subject, is in fact a very rare linguistic phenomenon in this corpus, being in one subcorpus

even completely absent. Therefore, it could not be used to make any kind of generalisation and it was excluded from the following analysis.

The intermodal perspective highlighted that the normalized frequency of linking adverbials was 273 in interpreted texts and 106 in translated texts, and as for apposition markers, 51 occurrences could be observed in the interpreted subcorpus and 30 in the translated one.

LINKING ADVERBIALS ITALIAN				
SUBCORPUS	**st-in-it**	**tt-in-it**	**st-tr-it**	**tt-tr-it**
RAW FREQUENCY	54	92	50	39
NORMALIZED FREQUENCY (100,000 words)	207	273	136	106
STATISTICAL COMPARISON significance indicated by: **(*)**	**monolingual comparable**		**monolingual intermodal**	
	st-in-it vs. tt-in-it	**st-tr-it vs. tt-tr-it**	**st-in-it vs. st-tr-it**	**tt-in-it vs. tt-tr-it**
***p*-Value** significant with $p < 0,05$	0.1125	**0.003396 (*)**	0.8448	**2.472×10^{-7} (*)**
APPOSITION MARKERS ITALIAN				
SUBCORPUS	**st-in-it**	**tt-in-it**	**st-tr-it**	**tt-tr-it**
RAW FREQUENCY	27	17	28	11
NORMALIZED FREQUENCY (100,000 words)	104	51	76	30
STATISTICAL COMPARISON significance indicated by: **(*)**	**monolingual comparable**		**monolingual intermodal**	
	st-in-it vs. tt-in-it	**st-tr-it vs. tt-tr-it**	**st-in-it vs. st-tr-it**	**tt-in-it vs. tt-tr-it**
***p*-Value** significant with $p < 0,05$	**0.02186 (*)**	**0.0001117 (*)**	0.8928	0.1884

CHE + CONJUGATED VERB

SUBCORPUS	**st-in-it**	**tt-in-it**	**st-tr-it**	**tt-tr-it**
RAW FREQUENCY	2/3	6/9	0	7/9
FREQUENCY IN PERCENT	67%	67%	0%	78%

TABLE 6: Results – Italian

ST: source texts
TT: target texts
IN: interpreted
TR: translated

IT: Italian
EN: English

ST-IN: transcripts of the original speech
TT-IN: transcript of the interpreter's rendition
ST-TR: minutes in the original language
TT-TR: translated minutes

3.3.2 Major findings

As was noted in the previous section, only two out of the three explicitness indicators chosen could be taken into account for the macroanalysis in Italian, since the third resulted as being too rare in the investigated corpus. Given the longer list of reporting verbs chosen for Italian (22 items in this category, as opposed to the 8 selected for English on the basis of Kajzer-Wietrzny 2012), this finding is surprising. Maybe a preliminary contrastive analysis could help ascertain whether this indicator is not representative for the Italian language or if this lack of results is due to the corpus size.

Table 7 gives an overview of the major explicitness tendencies detected. Overall, the data show an opposite trend compared to that highlighted in the English subcorpora. In both interpreted and translated subcorpora there is a statistically lower occurrence of the indicators chosen, which might be ascribed to a tendency towards a lower degree of explicitness. The only exception to this is represented by the increased frequency of linking adverbials in the interpreted subcorpora, though this was not confirmed by Fisher's statistical test. The data concerning apposition markers are more homogenous, since both interpreted and translated subcorpora feature fewer explicitness items than their source counterparts.

	SUBCORPUS	**NORMALIZED FREQUENCY**	**EXPLICITNESS IN TARGET SUBCORPUS**	**STATISTICAL SIGNIFICANCE**
LINKING ADVERBIALS ITALIAN	tt-in-it	273	HIGHER	NO
	st-in-it	207		
	tt-tr-it	106	**LOWER (*)**	**YES**
	st-tr-it	136		

APPOSITION MARKERS ITALIAN	tt-in-it	51	**LOWER (*)**	**YES**
	st-in-it	104		
	tt-tr-it	30	**LOWER (*)**	**YES**
	st-tr-it	76		

TABLE 7: Explicitness tendencies – comparable perspective (Italian)

The intermodal investigation carried out in the Italian subcorpora shows once again an opposite trend compared to that observed in the English subcorpora, as indicated in Table 8, where explicitness tendencies are illustrated from an intermodal perspective. Here the occurrences of the selected sets of items were higher in the interpreted texts, but only for linking adverbials is this tendency statistically confirmed.

	SUBCORPUS	NORMALIZED FREQUENCY	EXPLICITNESS	STATISTICAL SIGNIFICANCE
LINKING ADVERBIALS ITALIAN	tt-in-en	273	**HIGHER (*) in tt-in-it**	**YES**
	tt-tr-en	106		
	st-in-en	207	NO statistically significant difference between source subcorpora	
	st-tr-en	136		
APPOSITION MARKERS ITALIAN	tt-in-en	51	HIGHER in tt-in-it	NO
	tt-tr-en	30		
	st-in-en	104	NO statistically significant difference between source subcorpora	
	st-tr-en	76		

TABLE 8: Explicitness tendencies – intermodal perspective (Italian)

3.4 Parallel microanalysis

3.4.1 Introduction

After having discussed the major explicitness tendencies observed thanks to a twofold monolingual analysis following a comparable and an intermodal approach, in this section some of the findings so far obtained are tested, performing a parallel qualitative analysis in English and Italian – a way of further exploiting the huge potential of EPTIC. In particular, the source frequencies of the selected sets of items are compared and investigated through bilingual concordances. Needless to say that only linking adverbials and apposition markers are the subject of this analysis, since these are the only indicators that can be contrasted from an inter-linguistic viewpoint.

The assumption underlying this kind of analysis, combined with the previous one, is that this is an effective way to verify if the emergent tendencies of higher or lower explicitness are in fact the result of explicitation or implicitation processes. For the purposes of this paper, only two interesting cases are discussed, namely English linking adverbials and Italian apposition markers.

3.4.2 Parallel microanalysis of English linking adverbials

In the monolingual quantitative analysis of linking adverbials, the tendency of translated texts being more explicit than comparable English written texts emerged as statistically significant. From a parallel point of view, the data related to linking adverbials (Table 9) show a drop in the number of the occurrences of this indicator from 54 in Italian original speeches to 15 in their English interpreted counterparts.

	SUBCORPUS	RAW FREQUENCY
LINKING ADVERBIALS ENGLISH	st-in-it	54
	tt-in-en	15
	st-tr-it	50
	tt-tr-en	50

TABLE 9: linking adverbials - parallel perspective (English)

By looking at parallel concordances, it can be observed that English interpreters tend to leave out linking adverbials more frequently than Italian speakers, as shown in the following examples (1) and (2).

(1)

st-in-it <text_id 18>: Un precedente ripeto molto grave di censura che sembra più tipico della democrazia alla cu- turca che non di un paese sicuramente democratico e fondatore dell'Unione europea contro la forza politica che vuole autonomia e federalismo e quindi contro lo statalismo //

tt-in-en: It 's a kind of thing that one would expect more in Turkey than in a country which is supposed to be democratic and which is certainly a founding country of the European Union // Ehm we are in favour of federalism and against the state centralism [...]

st-tr-it: E' un precedente , ripeto , molto grave di censura - che sembra più tipico della democrazia alla turca che di un paese sicuramente democratico e fondatore dell' Unione europea - nei confronti di una forza politica che vuole autonomia e federalismo, quindi contro lo statalismo.

tt-tr-en: This seems more typical of Turkish style democracy than that of a country that is undoubtedly democratic and was a founding member of the European Union. I would reiterate that this is an extremely serious precedent concerning censorship of a political party that wants independence and federalism and is, therefore, against statism.

(2)

st-in-it <text_id 46>: Credo che la scelta dell'Unio- dell'Unione europea sia corretta quella di avere lunghi tempi non di fermarsi immediatamente con una politica dove i nodi vengono tagliati come nell'antico attraverso il taglio del nodo gordiano // Abbiamo bisogno quindi di un tempo abbiamo bisogno di riflettere e di costruire relazioni //

tt-in-en: I think that the European Union has made the right choice that is looking at things on the long-term rather than simply cutting the gordian knots immediately in the form of certain policies // What we need is time we need to reflect we need to build up our activities //

st-tr-it: Credo che sia corretta la scelta dell' Unione europea di optare per il lungo periodo e di non pretendere risultati immediati attraverso il taglio del nodo gordiano di antica memoria. Abbiamo quindi bisogno di tempo per riflettere e costruire relazioni.

tt-tr-en: I think the European Union 's decision to opt for the long term and not to aim for immediate results by cutting the Gordian knot of old is the right one. We therefore need time to reflect and to build relationships.

In example (1), being against state centralism is, in the speaker's version, a result of being federalist, while the interpreter simply juxtaposes the two concepts. Probably, it cannot be said that the interpreter's rendition lacks coherence, but certainly English recipients need to process the concept heard to a deeper level than the Italian audience. Example (2) shows a case in which a common Italian SVO pattern is turned by the interpreter into a pseudo-cleft sentence, placing more emphasis on the concept of time. Since this sort of structure presumably requires a more considerable cognitive effort for the interpreter, it might be that the interpreter opted for this more complex structure to compensate the linking adverbial's omission.

The observed tendency of English interpreters to omit linking adverbials that are, on the contrary, present in the respective translations, seems in line with the results of the intermodal analysis, thus confirming the higher explicitness of translated texts for this explicitness indicator.

3.4.3 Parallel microanalysis of Italian apposition markers

The quantitative comparable macroanalysis of Italian apposition markers highlighted a statistically significant tendency towards a lower degree of explicitness of the interpreted subcorpus compared to the original oral one. Table 10 shows the results of the investigations carried out on this indicator from a parallel perspective, displaying a rise from 11 to 17 occurrences in interpreted texts.

	SUBCORPUS	RAW FREQUENCY
APPOSITION MARKERS ITALIAN	st-in-en	11
	tt-in-it	17
	st-tr-en	9
	tt-tr-it	11

TABLE 10: apposition markers - parallel perspective (Italian)

By examining the relevant parallel concordances, the Italian interpreters' preference to add apposition markers can be clearly observed in examples (3) and (4).

(3)

st-in-en <text_id 23>: There is sometimes a danger that we engage in the in the politics of the Book of Genesis //

tt-in-it: a volte c'è un pericolo ossia ehm fare un po'come la Genesi che luce sia e luce fu //

st-tr-en: Sometimes there is a danger that we engage in the politics of the Book of Genesis : let there be light - and there is light.

tt-tr-it: A volte si rischia di fare la politica del Libro della Genesi : " Sia la luce. E la luce fu ".

(4)

st-in-en <text_id 3>: Commissioner Byrne I welcome very much your statement here here this morning // But I understand Sir you went to Thailand and were told that it wasn't avian flu but it was chicken cholera cholera //

tt-in-it: Commissario Byrne io sono molto lieto della dichiarazione che lei ha fatto questa mattina però mi pare che lei sia andato in Tailandia e le è stato detto che non c'era l'influenza aviaria ma che era un'altra malattia dei polli e cioè il colera dei polli //

st-tr-en: Mr President, Commissioner Byrne, I welcome your statement here this morning, but I understand that you went to Thailand and were told that it was not avian flu but chicken cholera.

tt-tr-it: Signor Presidente, Commissario Byrne, accolgo con favore la sua dichiarazione di stamani, ma ho sentito che lei è stato in Tailandia e che le è stato detto che non si trattava di influenza aviaria, ma di colera dei gallinacei.

While example (3) seems to be a simple addition of the apposition marker *ossia*, without further relevant interventions by the interpreter on his/her production, example (4) features a more segmented rendition than the original speech. The structure "it wasn't x, but y" is rendered into Italian more gradually "it wasn't x, but another illness, that is y", and is a typical case of the interpreter approaching the correct translation by stating a general

term first and then correcting and/or complementing it with a more precise one immediately afterwards. The choice of a longer rendition might have been required by the interpreter's need to recall the information from the short-term memory, or to think of a more accurate term. In both examples, apposition markers appear as good tools to add a new piece of information to an utterance that could also have been considered as concluded, without starting a new sentence.

Finally, example (5) shows a quite common phenomenon in the corpus. Here the apposition marker is added in the interpreted production as a consequence of the interpreter verbalising a logical link that s/he presumably perceives from the speaker's prosody. In the corresponding written versions, this link is not lexicalised but it is expressed through punctuation marks.

(5)

st-in-en <text_id 23>: Mention has been made here about the slow progress in relation to the two major ehm Amsterdam imperatives the ehm the ehm the directives in relation to asylum //

tt-in-it: Si è parlato della lentezza dell'avanzamento dei due imperativi più importanti di Amsterdam ossia le di- ehm le direttive legate all'asilo //

st-tr-en: Mention has been made of the slow progress in relation to the two major Amsterdam imperatives: the directives in relation to asylum.

tt-tr-it: Si è parlato dei lenti progressi riguardanti i due principali imperativi di Amsterdam**:** le direttive in materia di asilo.

In the light of the cases of explicitation examined so far, it seems reasonable to conclude that the tendency of interpreted texts being less explicit cannot be observed in the parallel qualitative analysis performed.

4. Conclusions

This study found no clear evidence of more or less explicitness in interpreted/translated versus untranslated speeches, and therefore no evidence for a universal tendency in its strictest sense. In other words, no homogenous and conclusive tendency could be observed which could summarise and include all the results obtained with the methodology applied to analyse each explicitness indicator selected. On the other hand, from the monolingual macroanalysis different tendencies according to language direction have emerged.

Firstly, results in the English subcorpora in this study suggest an overall tendency of increased explicitness in translated and interpreted subcorpora, but only the tendencies of the translated subcorpus being more explicit compared to the written comparable one and to the interpreted subcorpus were confirmed by Fisher's exact test of statistical relevance. The results highlighted by the queries related to both linking adverbials, apposition markers and optional *that* do not indicate a statistically significant tendency towards a higher explicitness of interpretations, while it seems that their frequency is, in a few cases, statistically significantly more pronounced in translated texts and not statistically significant in the remaining ones. The findings related to all the three explicitness indicators selected are therefore in line with Kajzer-Wietrzny's (2012: 141).

Secondly, after having excluded the third explicitness indicator because of its overall lower frequency in the corpus, an overall tendency of lower explicitness of both interpreted and translated texts has emerged from the investigations in the Italian subcorpora. As for linking adverbials, the tendency of increased explicitness of interpreted texts compared to source oral texts in Italian was not statistically significant. On the other hand, the tendency of translated texts being less explicit than their comparable counterparts was confirmed by Fisher's exact test. Also the lower frequency of apposition markers in both translated and interpreted texts compared to the respective comparable subcorpora was statistically proved. From an intermodal point of view, the tendency of interpreted texts being more explicit than translated ones was statistically relevant only for linking adverbials. It must be borne in mind that Kajzer-Wietrzny's study did not investigate Italian subcorpora, and therefore these results not only corroborate hers, but also complement them.

Finally, thanks to the versatility of the EPTIC corpus and the twofold methodology (quantitative and qualitative) applied, some interesting differences according to language direction have emerged from the parallel microanalysis. In spite of the limitations that are present in this study, such as the sample size and the unbalanced subcorpora, in this paper the English interpreters' preference to leave out linking adverbials more frequently than Italian speakers, and the tendency of Italian interpreters to add apposition markers, which they use more frequently than English speakers were discussed. In particular, the first case confirms the outcome of the monolingual analysis, while in the second case the findings of the quantitative macroanalysis were disproved by the parallel qualitative microanalysis.

The overriding aim of this study was to conduct an investigation into explicitness, trying to add new elements to the discussion on translation and interpreting universals. Hopefully, the methodology applied will aid in comparing and contrasting the findings of future studies, thus enriching the debate on translation and interpreting universals. Furthermore, parallel concordances allowed for the highlighting of some shortcomings of traditional monolingual comparable analyses, whose results could not sometimes be disconfirmed by parallel investigation. A deeper analysis on larger Italian corpora of the explicit structure after reporting verbs in clauses with the same subject is hoped to verify its appropriateness. We also noted (3.3.2) that a preliminary contrastive analysis can more effectively help choose perfectly equivalent explicitness indicators, especially if this kind of investigation is to be carried out on non cognate languages. These and other aspects may represent a fertile ground for future studies aimed at dispelling some of the confusion about this potential universal.

Nevertheless, these results highlight the huge untapped potential of bilingual, bidirectional and intermodal corpora like EPTIC, and the need to enlarge the corpus in order to gain further research insights into the nature of translation and interpreting universals.

References

Baayen, R. H. (2008) *Analyzing Linguistic Data*, Cambridge, Cambridge University Press.

Baker, M. (1993) "Corpus Linguistics and Translation Studies: Implications and Applications" in *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis and E. Tognini-Bonelli (eds) Amsterdam, John Benjamins, 233–50.

---- (1996) “Corpus-based Translation Studies: The Challenges that Lie Ahead” in *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, H. Somers (ed) Amsterdam, John Benjamins, 175–86.

Baumgarten, N., B. Meyer and D. Özçetin (2008) “Explicitness in Translation and Interpreting: A Critical Review and Some Empirical Evidence (of an Elusive Concept)”, *Across Languages and Cultures* 9, no.2: 177–203.

Becher, V. (2010a) “Towards a More Rigorous Treatment of the Explicitation Hypothesis in Translation Studies”, *Trans-kom* 3, no. 1: 1–25.

----(2010b) “Abandoning the Notion of "Translation-inherent" Explicitation: Against a Dogma of Translation Studies”, *Across Languages and Cultures* 11, no. 1: 1–28.

Bendazzoli, C. (2010) “The European Parliament as a Source of Material for Research into Simultaneous Interpreting: Advantages and Limitations” in *Translationswissenschaft - Stand und Perspektiven*, L. N. Zybatow (ed) Frankfurt, Peter Lang: 51–68.

Bernardini, S., A. Ferraresi and M. Miličević (2013) *From EPIC to EPTIC - building and using an intermodal corpus of translated and interpreted texts*. Paper presented at the *46th Annual Meeting of the Societas Linguistica Europea (SLE 2013)* Split, Croatia. 18-21 September.

Blum-Kulka, S. (1986/2001) “Shifts in Cohesion and Coherence in Translation” in *The Translation Studies Reader*, L. Venuti (ed) (2001) London, Routledge: 298–313.

Chesterman, A. (2004) “Beyond the Particular”, in *Translation Universals. Do They Exist?*, A. Mauranen and P. Kujamäki (eds), Amsterdam, John Benjamins: 33–49.

Gumul, E. (2006) “Explicitation in Simultaneous Interpreting: A Strategy or a By-product of Language Mediation?”, *Across Languages and Cultures* 7, no. 2: 171-90.

Ishikawa, L. (1999) “Cognitive Explicitation in Simultaneous Interpreting”, in *Anovar/anosar estudios de traducción e interpretación,* A. Àlvarez Lugrís and A. Fernándey Ocampo (eds), Vigo, Universidade de Vigo: 231–57.

Kajzer-Wietrzny, M. (2012) *Interpreting Universals and Interpreting Style*, PhD diss., Adam Mickiewicz University, Poznań.

Konšalová, P. (2007) “Explicitation as a Universal in Syntactic Ce/condensation”, *Across Languages and Cultures* 8 no. 1: 17-32.

Mauranen, A., and P. Kujamäki (eds) (2004) *Translation Universals. Do They Exist?*, Amsterdam, John Benjamins.

Olohan, M., and M. Baker (2000) “Reporting That in Translated English. Evidence for Subconscious Processes of Explicitation?”, *Across Languages and Cultures* 1, no. 2: 141–58.

Øverås, L. (1998) “In Search of the Third Code: An Investigation of Norms in Literary Translation”, *Meta* 43, no. 4: 571–88.

Pápai, V. (2004) “Explicitation. A Universal of Translated Text?” in *Translation Universals. Do They Exist?*, A. Mauranen and P. Kujamäki (eds), Amsterdam, John Benjamins: 143–64.

Puurtinen, T. (2004) “Clause Connectives in Finnish Children's Literature” in *Translation Universals. Do They Exist?*, A. Mauranen and P. Kujamäki (eds), Amsterdam, John Benjamins: 165–75.

Pym, A. (2005) “Explaining Explicitation”, in *New trends in Translation Studies. In Honour of Kinga Klaudy*, K. Károly and A. Fóris (eds), Budapest, Akadémiai Kiadó: 29–34.

Sandrelli, A. and C. Bendazzoli (2005) “Lexical Patterns in Simultaneous Interpreting: A Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)”, *Proceedings from the Corpus Linguistics Conference Series* 1, no. 1.

Schjoldager, A. (1995/2002) “An Explanatory Study of Translational Norms in Simultaneous Interpreting” in *The Interpreting Studies Reader*, F. Pöchhacker and M. Shlesinger (eds) (2002), London & New York, Routledge: 301-11.

Serianni, L. (1989). *Grammatica Italiana*, Torino, Utet.

Shlesinger, M. (1995) “Shifts in Cohesion and Simultaneous Interpreting”, *The Translator* 1, no. 2: 193–214.

Notes

[1] S-Universals are ‘universal differences between translations and their source texts’, while T-Universals are ‘universal differences between translations and comparable non-translated texts’ (Chesterman 2004).

[2] Part-of-speech tagging and lemmatization was performed independently of EPIC using Tree-Tagger, while Corpus Work Bench (CWB) was used for the indexing process (Bernardini et al. 2013).

[3] In this study, by ‘quantitative analysis’ we mean the analysis of the number of occurrences of the indicators chosen, on which we applied Fisher’s exact test to find out their statistical significance.

[4] For further information about the *R project* see: **http://www.r-project.org/** (accessed: 16 June 2016).

[5] This is also the reason why chi-squared test was not applied. In this respect, Baayen maintains that ‘For tables with not too large counts, a test of independence of rows (or columns) that produces more precise *p*-values is Fisher’s exact test’ (Baayen 2008: 113).

[6] For the results of the different indicators and corpora to be comparable, raw frequencies have all been normalized per 100,000 words. Differently, the occurrences of the third indicator of each set of indicators are already expressed in percentages, and hence comparable across subcorpora.

[7] P-Values with more than five decimal numbers are conveniently written in scientific notation.

"Interpreting Universals: A study of explicitness in the intermodal EPTIC corpus", *inTRAlinea* Special Issue: New Findings in Corpus-based

Interpreting Studies.

Stable URL: http://www.intralinea.org/specials/article/2320

"Looking up phrasal verbs in small corpora of interpreting An attempt to draw out aspects of interpreted language", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2319

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy. This printout was generated directly from the online version of this article and can be freely distributed under Creative Commons License **CC BY-NC-ND 4.0**.

Looking up phrasal verbs in small corpora of interpreting
An attempt to draw out aspects of interpreted language

By Andy Cresswell (University of Bologna, Italy)

Abstract & Keywords
English:

A comparison of the frequencies of English phrasal verbs in two corpora of simultaneous interpreting into English, one B into A, and the other A into B, showed the A to B interpreters used phrasal verbs much less often, which confirms the theories that non-natives find it hard to acquire phrasal verbs, and that non native interpreters, because of the pressure of the SI process, find it hard to access them. In particular, phrasal verb lemmas with idiomatic and figurative meanings were very much less frequent in the language of the A into B interpreters, as were phrasal verbs with aspectual meaning. Despite this, metafunctional analysis showed that the A into B interpreters did use phrasal verbs with interpersonal function, to pursue the role of interpreter as mediator. The most striking finding on metafunctions was that phrasal verbs used by the B into A interpreters for the process of textualisation, and which are therefore crucial for meaning assembly, were almost entirely absent in the language of the A into B interpreters. In short, phrasal verbs are clearly an important resource for SI and the lack of them in the language of A into B interpreters suggests an urgent instructional need.

Keywords: simultaneous interpreting, directionality, collocations, phrasal verbs, interpreter training

1. The scope of this article

The primary purpose of this article is to make the case for the teaching of English phrasal verbs in language lessons for trainee interpreters. Phrasal verbs are a major sub-component of formulaic language, and the use of formulaic language is a major factor in the fluency that makes simultaneous interpreting a feasible activity. Despite this, to the best of my knowledge, nothing has so far been published that focuses specifically on the role played by phrasal verbs in simultaneous interpreting.

The article assumes that simultaneous interpreting is a process of assembly of meaning, as illustrated by Setton (1998/2002), and that it often necessarily involves deletion and summarising of aspects of the source text. For these reasons the article deliberately does not take a "translation" oriented perspective, and is not concerned with observing the items of language that phrasal verbs are translating, but rather with the more general questions of whether phrasal verbs are used by native and native-like interpreters, and whether there is therefore a need for non-native interpreters to use them. In seeking to find this out, the methods used are primarily those of counting and the comparison of frequencies in corpora. Given the lack of availability of transcribed interpretations, the corpora are necessarily small, but Aston (1997) argues that interesting results and applications can be derived from small corpora, and Flowerdew (2004) points out that if a corpus is specialised enough, small size is enough to provide satisfactory results – albeit when the corpora are very small, as in this research, an appropriate degree of caution will need to be exercised when generalising.

Assuming that phrasal verbs are used by interpreters, the article aims to discover which phrasal verbs are particularly characteristic of interpreting in comparison with the English language as a whole, as represented by the British National Corpus. Finally, by studying the context of use as represented in concordance lines, the article seeks to show what functions interpreters use phrasal verbs for, with a view to possible use in a language syllabus for trainee interpreters.

2. Directionality, Proficiency and Formulaic Sequences

The cognitive complexity of the process of simultaneous interpretation requires extensive knowledge of the source and target languages, while time constraints imply there must be swift access to that knowledge both in the sense of plausibly accurate comprehension and in the sense of adequately synchronic production. The initial response to this challenge was that interpreters should work only towards their mother tongue (Herbert 1952: 61), and that simultaneous interpreting, in particular, should ideally be the exclusive preserve of native speakers (Seleskovitch 1978: 100), working from their foreign active or passive language(s) (or B and C languages, respectively) to their native language (or language A according to AIIC's classification). This favouring of C/B to A directionality is reflected in the position taken by AIIC (as reported by Bartłomiejczyk 2004: 247). Yet there are plausible arguments on the other side. Denissenko (1989: 157) argues that A to B directionality, which implies the ability to comprehend of the native speaker, but the more limited ability to produce language of the non-native speaker, is more likely to lead to reproduction of all or most of the original message. In addition, there is the pragmatic argument, as highlighted by Bartłomiejczyk (2004: 247), that circumstances lead to market demand for A to B in addition to retour, and that interpreting students might as well be properly prepared for this. One essential aspect of such preparation for working towards B, in other words for working towards the non-native language, is the effort to bring the productive language skills of the non-native speaker nearer to the proficiency level of the native speaker.

So what is it about native speakers that non-native interpreters, in seeking to be native-like, should seek to emulate? From the points of view both of comprehension and of production, acquiring the phraseological knowledge of the native-speaker would appear to be both a feasible and a useful strategy for reducing stress and processing load in SI – especially for trainees, since, according to Setton (1998/2002: 199), the attention taken by selective suppression,

working in two languages, and by meaning assembly, makes access to phrasal expressions the only automatic mechanism that is likely to be available. From the point of view of comprehension, such phraseological knowledge is an important component of pragmatic processing in oral comprehension (Rost 2011: 138). From the point of view of production, the fact that native speakers know formulaic sequences (Wray 2002), in other words ready-formed phrases or strings or slot and filler patterns, brings processing advantages in reducing cognitive load and freeing up attention. This is because formulaic sequences are retrieved from memory as whole units – as it were, automatically, so that the time and effort needed to encode meanings through combining grammar and lexis is not needed (Pawley & Syder 1983: 192; Wray 2002: 9). In the words of Pawley and Syder (2000: 164), 'It is knowledge of conventional expressions, more than anything, [...] that is the key to nativelike fluency'.

3. Phrasal verbs: description and use

3.1 The value of phrasal verbs

Multiword verbs, together with their collocations, constitute a major subclass of formulaic expressions. Following Biber et al. (1999), the subclass can be further divided, on the one hand, into free combinations of verbs and prepositions, which are not formulaic sequences, and on the other, into prepositional verbs, phrasal verbs, phrasal-prepositional verbs, and other multiword verb constructions, which are formulaic sequences (Biber et al: 1999: 403–427). Among these, the principal category is phrasal verbs, which are formed of a lexical verb in combination with an adverbial particle.

Many phrasal verbs are polysemic – they are quasi-empty vessels, with a vague tendency towards meaning which ends up being defined by the words used around them. As a consequence, many of these verbs can be used in quite a wide range of contexts – and the cotexts (in other words the adjacent text with which the phrasal verb is closely linked semantically, such as its regular collocates) may themselves extend the phrasal verb into longer strings.

As phrasal verbs are often short, with monosyllabic lexical verbs and particles, they also arguably offer value to the interpreter in reducing articulation time and effort. Additionally, they are fairly frequent in English, occurring in the British National Corpus (henceforth the BNC) at a rate of approximately one every 192 words (Gardner and Davies 2007: 347). For reasons then of processing advantages and expediency, combined with their high frequency in the English language as predicted by the BNC, one would expect to find a reasonably high occurrence of phrasal verbs in English produced by English native-speakers and native-like simultaneous interpreters.

3.2 Phrasal verbs: structural and semantic processing difficulties for non-native interpreters

The difficulty non-native speakers experience in acquiring English phrasal verbs is well documented in the literature (Dagut and Laufer 1985; Laufer and Eliasson 1993; Celce-Murcia and Larsen-Freeman 1999: 425; Liao and Fukuya 2004; Siyanova and Schmitt 2007). The difficulty is experienced even by speakers of Germanic languages, such as Dutch, which themselves have phrasal verbs (Hulstijn and Marchena1989). While one would expect experienced native and native-like interpreters working towards English to have moved beyond such difficulties, this is not necessarily the case where interpreters are non-natives working towards English as a B language, particularly if they are trainees or beginning professionals. The difficulties can be divided into two types. The first type of difficulty is structural, and applies particularly to speakers of languages with few or no phrasal verbs. The problem here is that, given the pressure experienced during SI, as detailed by Setton (1998/2002: 199; see section 2 above), interpreters working towards B may retreat to more automatised structures analogous to those of their mother tongue, avoiding the use of phrasal verbs. Given that phrasal verbs are so difficult to acquire, and that even in non-interpreters of advanced proficiency levels there is a tendency to avoid using them (Siyanova and Schmitt 2007: 129), one would expect that interpreters will to some extent avoid using them when working towards English as a B language in the booth.

The second major factor determining the difficulties of English phrasal verbs for non-natives is semantic. The non-literal nature of the meaning that many phrasal verbs convey arguably creates processing problems during construal of the message. To be clear, I am using the term "literal" with the sense intended by Grant and Bauer (2004: 39), who quote Lakoff's definition of literal as 'nonmetaphorical literality: directly meaningful language – not language that is understood, even partly, in terms of something else' (Lakoff 1986: 292). According to Celce-Murcia and Larsen-Freeman (1999), cited in Darwin and Gray (1999: 68), phrasal verbs can be divided into three categories on the basis of degree of literalness: literal, idiomatic, and aspectual. Proceeding according to Lakoff's definition, in Celce-Murcia's first category, a literal phrasal verb is exemplified by *take* and *down* in *take down the poster.* The second category, idiomatic phrasal verbs, is exemplified by *make up* with the meaning of 'become reconciled', where (to use the analytical method proposed by Grant and Bauer 2004: 44, which relies on Frege's principle of compositionality as cited in Lyons 1995: 24) the meaning of the phrase is not recoverable from any dictionary definition of the word *make* combined with any dictionary definition of the word *up* – in other words, the meaning of *make up/(=become reconciled)* is non-compositional. Alongside such non-compositional items as *make up*, which Grant and Bauer would call 'core idioms', there are phrasal verbs whose meanings are recoverable by means of a shared understanding, between utterance producer and utterance recipient, of figures of speech such as metaphor (Grant and Bauer 2004: 49). One example is *stand out*, as in *his writing stands out among that of his contemporaries*, but while this example is metaphorical, Grant and Bauer (2004: 49) point out that figurative language (obviously including figurative phrasal verbs) includes all figures of speech, whose meanings are all equally recoverable through 'taking a conversational untruth and extracting probable truth from it by an act of pragmatic interpretation' (Grant and Bauer 2004: 50). This quotation can be taken as a definition of figurative language.

It should be stated here that most writers on phrasal verbs, and on idiomaticity in general, do not distinguish between idiomatic meaning and figurative meaning. In Grant and Bauer's view (Grant and Bauer 2004) there are three categories of "idiomatic" meaning – first, core idioms, which are completely non-compositional, second, phrasal items with one element that is non-compositional, and third, figurative language. But it appears from Siyanova-Chanturia and Martinez (2015), who review the literature on processing of non-literal phrases, that in this field of research, the three categories are usually conflated. Hence in the context of the current research on interpreted language with its concomitant processing constraints, I will refer to "idiomatic and figurative" phrasal verbs.

Celce-Murcia's third category of aspectual phrasal verbs in not unproblematic either, in involving overlap with the conflated idiom/figurative category. This is the definition of aspectual phrasal verbs in the words of Darwin and Gray (1999: 68):

> Whereas the verb proper in aspectual phrasal verbs can be understood literally, the particle contributes meanings, not commonly understood, about the verb's aspect. For example, *up* in *They ate up all the chips and drank up all the soda* signals that the actions are complete.

Aspectual verbs are thoroughly discussed by Side (1990), but many of his examples are figurative in addition to being aspectual, for example *took off* in *his business really took off* (Side 1990: 148) derives its meaning not only from the aspectual *off* meaning departure, but also by analogy with an aircraft taking off.

It is worth commenting on Darwin and Gray's phrase 'not commonly understood', for if a non-native speaker is unaware of the aspectual meaning of particles, aspectual phrasal verbs become in effect non-compositional. On the other hand, in the examples quoted by Darwin and Gray (1999, above), for comprehension purposes the aspectual particle does not need to be understood as the meaning is easily recoverable in the context from the lexical verbs *ate* and *drank*. In these examples the potential problems for non-native speakers (henceforth, NNS) deriving from lack of knowledge of the aspectual meaning of *up* would lie in production, in the inability to produce the pragmatic effect of completion.

Whatever the semantic classification as far as non-literal meaning goes, there is general agreement that non-literal phrasal verbs cause additional difficulty to NNS. Siyanova and Schmitt (2007: 132) cite Laufer (1997), Moon (1997) and Wray (2000) as finding that 'both teachers and learners find idiomatic multi-word units more difficult than their nonidiomatic counterparts, which is likely to lead to avoidance behaviour' (Siyanova and Schmitt 2007: 132). A reading of Siyanova-Chanturia and Martinez (2015: 555) demonstrates that these difficulties have been observed empirically, and while the cited studies did not focus specifically on phrasal verbs, their conclusions about idiomatic phrases in general can be inferred to apply to idiomatic/figurative phrasal verbs, which are part of the same category. Hence, out of four studies comparing idiom processing by native speakers (henceforth NS) and NNS, three (Cieslika 2006; Underwood, Schmitt and Galpin 2004; Siyanova-Chanturia, Conklin and Schmitt 2011) reported NNS as processing literal uses of words used in idioms more quickly than they processed the same words used figuratively or idiomatically, while the reverse held for NS, who were shown to gain processing advantages from the use of idiomatic phrases (Siyanova-Chanturia and Martinez 2015: 555).

Since phrasal verbs form an integral part of the formulaic expressions that facilitate fluent production in English, it is worthwhile highlighting them for interpreter trainees and on inservice professional development courses. The best way to do this is to identify and describe situated use of phrasal verbs in interpreted language, produced during SI by experienced professional interpreters. This can be done by identifying the phrasal verbs in a corpus of interpretations, so that they can be presented in their context in materials used in language courses that supplement courses of training of conference interpreters.

3.3 ELF and the survival of the international phrasal verb

A further point concerning phrasal verbs is the issue of what has become known as ELF or English as a Lingua Franca (Jenkins 2001; McArthur 2003; Modiano 2003; Seidlhofer 2007), involving the simplification of English to exclude those features that non-natives find hard to use. If English in Europe is increasingly used as a medium of communication between non-native speakers, and if phrasal verbs are inherently difficult to acquire, does that mean that they are falling out of use in international forums? And that interpreters will therefore use them less? There may be a certain precedent for this in frequencies of phrasal verbs in some ex-colonial varieties of English, such as Indian English, which Schneider, working on ICE (the International Corpus of English) found to be rather low (Schneider 2004: 235). On the other hand, in Singapore, where English is a second language for nearly all of the population, yet is the national language and therefore universally taught to a reasonably high standard, the frequency of phrasal verbs is higher than in British English (Schneider 2004: 235). Additionally, research shows that phrasal verbs are relatively frequent in the international context represented by written EU documents. A study using CEUE (the Corpus of EU English), a 200,000 word corpus of EU documents intended for the general public, found a frequency of 1 every 200 words (Trebits 2009: 276), which is comparable to the frequency of 1 every 192 words found in the BNC (Gardner and Davies 2007: 347). These arguments suggest that, in international contexts where high levels of proficiency are a priority, phrasal verbs not only survive, but thrive. And this in turn suggests that the high level of proficiency expected of interpreters working towards English will guarantee a reasonably high frequency of phrasal verbs in interpreted English.

3.4 Phrasal verbs in interpreted English – hypotheses and questions

Hypothesis 1. On the basis of the arguments in sections 3.1–3.3 above, in other words on the basis of (a) the processing advantages of multi-word expressions of which category phrasal verbs form a part (sections 2 and 3.1), of (b) their expediency in the sense that they are short and quick to say (section 3.1), of (c) their high frequency in the English language as shown by the BNC, and of (d) the finding that they occur frequently in international English as shown by Trebits (2009) (section 3.3), it was hypothesised that phrasal verbs will occur frequently in the English of native speaker and native-like interpreters.

Hypothesis 2. On the basis (a) that non-native speakers find it difficult to acquire phrasal verbs, as shown in section 3.2 above, and (b) of the likelihood that non-native interpreters will respond to the pressure experienced during SI by retreating to more automatised structures analogous to those of their mother tongue (also in section 3.2), it was hypothesised that phrasal verbs will not occur reasonably frequently in the English of non-native interpreters working towards English as a B language. The meaning of "reasonably frequently" is clarified in the next paragraph.

To operationalise these hypotheses, the frequencies of phrasal verbs were measured in two corpora. Hypothesis 1 was tested by measuring the frequency of phrasal verbs in a corpus of English language interpretations produced by interpreters of native-speaker standard, which was called INT-A. Hypothesis 2 was tested by measuring the frequency of phrasal verbs in a corpus of English language interpretations produced by non-native interpreters, which was called INT-B. The notions of "reasonably frequent" and "reasonably frequently" were operationalised by comparison with frequency in the language as a whole, as represented by the BNC. On this basis, if both Hypothesis 1 and Hypothesis 2 were confirmed, then the frequency of phrasal verbs in INT-A would be fairly close to the frequency of phrasal verbs in the BNC, while the frequency of phrasal verbs in INT-B would be much lower than the frequency of phrasal verbs in both INT-A and the BNC. This would be regardless of genres and written/spoken production, on the principle that the structural and semantic difficulties outlined in section 3.2 above are intrinsic to the phrasal verb in itself, independently of the genre or mode (spoken or written) in which it is used.

With the aim of providing descriptions of situated use that could be used in language lessons for trainee interpreters, as mentioned in section 3.2 above, it was decided (in addition to investigating Hypotheses 1 and 2) to use the study in an exploratory way, to find out precisely which phrasal verbs are shown by the data to be used in interpreted English, and for which functions they were used. To help structure an exploration of exactly how the phrasal verbs were used, and to permit evaluation of the role that phrasal verbs might play in furthering the task of the interpreter, reference was made to the concept of the three metafunctions in Hallidayan systemic functional grammar – interpersonal, textual and ideational (Halliday and Matthiessen 2014: 30).

The three metafunctions in Hallidayan linguistics – ideational, interactional and textual – are 'dimensions of language' (Halliday and Matthiessen 2014: 30), and as such, they are complementary, and they co-occur, as is clear from the quotations from Halliday and Matthiessen that I will now cite. Halliday and Matthiessen (2014: 30) refer respectively to the ideational and interactional metafunctions as follows:

> every message is both about something and addressing someone, and [...] these two motifs can be freely combined – by and large, they do not constrain each other.

They describe the role of the textual metafunction as follows:

> this can be regarded as an enabling or facilitating function, since both the others – construing experience and enacting interpersonal relations – depend on being able to build up sequences of discourse, organising the discursive flow, and creating cohesion and continuity as it moves along. This, too, appears as a clearly delineated motif within the grammar (Halliday and Matthiessen 2014: 30).

Complementarity and co-occurrence mean that two or three of the metafunctions can be realised in the same clause. This multifunctionality has an implication for processes of language analysis in Applied Linguistics. Items of contextualised language – words or phrases – should not be assumed to be exclusively associated with any single metafunction, on the principle of mutual exclusivity of categories. The contribution any linguistic item makes to the meaning is dependent on the other linguistic items it is associated with and on the context, and vice versa. For example, Hyland (2005: 59) cites Hood and Forey (1999), who show that in the context of shifting from a positive to a negative judgement, *but* and *however* realise interactional function, in addition to the more predictable textual function. Conversely, the same words and phrases used in different structures and different contexts can represent different metafunctions, as Martin (1992:8) shows in a worked example. Hence any judgements made in the current research about the predominant metafunction expressed by a given phrasal verb reflect the researcher's view of the role of that phrasal verb *in that context.* The researcher's judgement of the ideational, textual or interpersonal metafunction of each phrasal verb examined is necessarily subjective, although this subjectivity was constrained by repeated analysis after a month's interval, which resulted in an index of consistency of 98 per cent.

One reason here for distinguishing in the data between the three metafunctions is that it permits evaluation of the potential value of the effort taken to learn a particular phrasal verb, through distinguishing between interpreting process, which is predictable, because it recurs, and interpreted speaker's translated product, which is unpredictable. I will argue that phrasal verbs with textual and interpersonal metafunction are used as part of the interpreting process, and that there is a good probability that the contexts where they are used will recur; these verbs are worth practising to make them automatically retrievable. On the other hand, with phrasal verbs with the ideational metafunction, the probability of their recurrence is a function of the extent to which the interpreter is or is not always dealing with the same subject matter – and for interpreters of both parliamentary debates and of conferences, subject matter is inherently variable. This distinction is subject to the qualification that some aspects of ideational metafunction, such as some mental processes or some material processes, are likely to recur in repeated contexts and are likely to prove productive.

The need for vocabulary to express textual metafunction is somewhat more predictable. As Setton describes it, the re-creation of textual structure is part and parcel of the process of simultaneous interpretation:

> relevance (coherence) is sought as a matter of routine. Context is seen as a nested set of assumptions: a background about the world and the situation, then assumptions based on previous discourse, on the previous utterance, and finally those taking shape from the utterance-initial cues about the illocution or modality of the ongoing argument [...]" (Setton 1998/2002: 193).

So it is clearly worthwhile identifying phrasal verbs used with textual metafunction, describing their exact functions, and presenting them to trainees as high priority lexical items.

Similarly, the interpreter's function as mediator implies that any phrasal verbs used with interpersonal metafunction are likely to prove useful on repeated occasions; as with phrasal verbs used with textual metafunction, it is worth describing the exact functions to present to trainees. Hence identifying metafunctions points the way to phrasal verbs used in the interpreting process, which constitute high priority lexical items for inclusion on a language syllabus for trainee interpreters.

On the basis of the probability that they are likely to be experienced at textualisation towards English, and at the same time are likely to take advantage of the processing advantages of use of phrasal verbs as outlined in sections 2 and 3.1, it was predicted that the native or native-like interpreters of INT-A would tend to use phrasal verbs for textual metafunction. On the basis of the difficulties that non-natives have in producing phrasal verbs as outlined above, and on the basis that non-native interpreters have relatively little experience of textualisation towards English, it was predicted that non-native interpreters working towards English as B would tend not to use phrasal verbs with textual metafunction. Expressed as a hypothesis, this is

Hypothesis 3. Native or native-like interpreters will use phrasal verbs with textual metafunction more frequently than non-native interpreters.

On a similar basis of the likely experience of the native or native-like interpreters working B to A of mediating aspects of interpretation, combined with their knowledge of ready-made phrasal verb expressions on the one hand, and on the other hand on the basis of the difficulties non-natives have in producing phrasal verbs combined with the added processing load of the A to B interpreter, the following hypothesis was formulated.

Hypothesis 4. Native or native-like interpreters will use phrasal verbs with interpersonal metafunction more frequently than non-native interpreters.

The final Hypothesis relates to the semantic processing difficulties highlighted in section 3.2.

Hypothesis 5. Given the evidence of processing difficulties experienced by NNS with idiomatic/figurative phrases, and of processing advantages experienced by NS, as cited in section 3.2 above, it can be hypothesised that non-native interpreters working towards their B language will use idiomatic/figurative phrasal verbs less often per 1000 words than native or native-like interpreters.

4. The Corpus Data

Two corpora of interpreted English, which I called INT-A and INT-B, were compiled out of several pre-existing corpora. The objective of INT-A was to represent the interpreted English of established professional interpreters, who should be either native speakers or native-like. It was made up of texts from three sources. The first two were corpora based on proceedings of the European Parliament, and consisted of interpreting towards English (from Italian and Spanish) taken from the EPIC corpus (Monti et al. 2005; Sandrelli, Bendazzoli and Russo 2010), plus all the interpreting towards English (from unspecified languages) in 2249, a corpus of all the English spoken in the European parliament on a single day (see Aston 2017), the non-interpreted sections from which had been removed on the basis of the tags. The third source was the DIRSI corpus of conference interpretations (Bendazzoli 2010), which contributed all of its native-speaker-interpreted English, all of which took place at two conferences about cystic fibrosis (Bendazzoli 2010: 155). The interpreters in the texts contributed by EPIC and DIRSI were tagged as native speakers, but as there was no information about the interpreters in 2249, it was assumed that they were (a) established professionals and (b) either native speakers or native-like – and a reading of the interpretations in 2249 to evaluate their fluency and accuracy provided no evidence to counter this supposition. The total number of running words in INT-A was 85,891.

The INT-B corpus was intended to represent NNS interpreter language. It was made up of all those transcripts of interpretations towards English in both the DIRSI corpus and the EPIC corpus that were tagged as being done by non-native interpreters. All the texts from DIRSI were translated into English from Italian, and were taken from the same two conferences about cystic fibrosis that supplied the native speaker material mentioned above, plus a conference about social care policy (Bendazzoli 2010: 155). The total number of running words in INT-B was 24,122. It is regrettable that INT-B was so much smaller than INT-A, but this reflects the easier availability of A-direction transcripts.

The transcription of all the words in all the component texts of both INT-A and INT-B was orthographic, and word for word, meaning that features such as reformulation were maintained. In 2249, punctuation was added, as it would be in writing. On the other hand, the transcription of all the texts taken from EPIC and DIRSI was without punctuation, with the speech divided into "units of meaning" using double slash marks (for further details, see Monti et al. 2005, section 2; Bendazzoli 2010: 184ff).

5. Interrogating the Data

Information relative to the research questions was uncovered by making KWIC concordances of phrasal verbs using *Wordsmith Tools Concord*, first version 6, and later version 7 – see Scott (2012/16), and by comparing the frequency data obtained with the data about phrasal verbs in the BNC in Gardner and Davies (2007). There are various definitions of the term "phrasal verbs" (for a review, see Darwin and Gray 1999), but for the sake of simplicity I followed the definition of the term in Gardner and Davies (2007), which is a simplified version of the exhaustive definition in Biber et al. 1999: 405ff). To paraphrase Gardner and Davies (2007: 344), my definition of a phrasal verb is a lexical verb followed by an adverbial particle, with meaning derived from a combination of the two.

The phrasal verb forms were retrieved from the data by searches for each of the following particles, which were those used in the searches of the BNC by Gardner and Davies (2007):

out/up/down/back/off/round/along/over/around/on/through/about/in/under/by/across.

The searches were carried out first in INT-A, then in INT-B. Each concordance line was then examined, first to eliminate instances of verbs and prepositions in free combination, and secondly, to eliminate prepositional verbs (for these two categories, see Biber et al. 1999: 405ff).

To verify whether multiword verbs in concordances were phrasal (and hence eligible for inclusion in the results) or prepositional (and thus excluded from the results) I used the following operational procedures, based on syntactic descriptions from Biber et al. (1999: 405ff):

A. is the verb+particle unit used transitively or intransitively? If – as in example (1) – the answer is "intransitively" (i.e. if the following noun answers the question *where* or *when*), then the unit is a phrasal verb, the particle being clearly adverbial, since prepositions need to be followed by noun phrase objects that answer the questions *who* or *what* – as in (2).

(1) I think the law will *go through*. Intransitive – so adverbial.

(2) We should *go through* these results before the presentation. Transitive use.

B. If the verb+particle unit is used transitively, can it be used with a pronoun object placed between the lexical verb and the particle? If the answer is no, as in (3), the unit is a prepositional verb and is excluded from the results.

(3) We should *go* them *through* X

If the answer is yes, as in (4), the unit is counted as a phrasal verb, and included in the results.

(4) It was only the whiskey that *saw* him *through*.

Other points on the selection of which forms were phrasal verbs are that as only lexical verbs are included in the definition, combinations with *be* were excluded, for example, *be out* (= be in the public domain). On phrase length, which is relevant because phrasal verbs can have noun phrase objects between verb and particle, only 2–4 word phrases were considered, following Gardner and Davies (2007: 345). Gerunds, for example *by setting up*, were classified as verb forms not noun forms, as they are in the BNC. Phrasal-prepositional verbs were counted as phrasal verbs, because they follow the Gardner and Davies (2007: 344) definition of a verb followed by an adverbial particle. Thus *come up with* was counted as the phrasal verb *come up*.

Once the above syntactic criteria were fulfilled, a phrase was accepted as a phrasal verb independently of any notional lists of canonical phrasal verbs, on the principle that phrasal verbs are an open-ended, productive category

in which the standard adverbial particles can combine with new lexical verbs to create new meanings (Side 1990: 146).

Once it had been verified that what remained in the concordances were phrasal verb forms only, the frequencies of each form were counted and entered on a spreadsheet. The frequencies were subsequently additionally grouped by lemma, to facilitate comparison with the frequencies in the BNC presented by Gardner and Davies (2007), which were taken to represent frequencies in the language as a whole. The lemmas were also ranked by frequency. For the metafunctional analysis, the concordances were merged and then classified according to the three Hallidayan metafunctions mentioned above.

When it was necessary to find the frequency in the BNC of a phrasal verb that was not listed by Gardner and Davies (2007), the online BYU-BNC (Davies 2004) was consulted, with a collocation search starting with the POS-tagged adverbial particle and looking for the lexical verb within a span of three words to the left. In this way, the phrases containing phrasal verbs found were two to four word phrases as they were in Gardner and Davies (2007).

For Hypothesis 5, phrasal verbs were classified as idiomatic/figurative if the meaning of the whole phrase was figurative, or if the meaning of the whole phrase was non-compositional in the sense that the meaning could not be deduced from summing the literal meanings (as defined in section 3.2 above) of the verb and the particle. If the meaning was entirely literal, they were classified as literal. With aspectual phrasal verbs, if figurative/idiomatic they were classified as figurative/idiomatic, and were classified as aspectual only if the meaning of the verb was literal, with aspectual use of the particle. The final test for literal meaning was any of the meanings of the individual word (verb or particle) listed in a contemporary dictionary (Cobuild 2001), according to Grant and Bauer's (2004) method as outlined in section 3.2 above. To help ensure plausibility of the classification, the analysis was repeated after a month's interval, with an index of agreement of 0.97.

6. Results and discussion

The INT-A corpus was used to make observations about the comparative frequency of phrasal verbs in interpreted language, while both corpora were used for the functional analysis and the analysis of literality. But before proceeding to report the results, I would like to clarify the statistical approach I have adopted to these comparisons.

When using mathematics in research, it is as well to consult a well-informed source for the most appropriate method. In corpus linguistics research, Kilgarriff, author of a review of statistical tests for corpus comparison (Kilgarriff 2001) published in *The International Journal of Corpus Linguistics*, constitutes such a source. As Kilgarriff has not only pointed out but also demonstrated empirically (in Kilgarriff 2005), in corpus frequency comparison the fact that language is a matter of choice, and is therefore not random, presents problems for statistical hypothesis testing (Kilgarriff 2005: 264). Hypothesis testing depends on disproof of the null hypothesis that there is only a random relationship between two sets of data (Kilgarriff 2005: 263). Clearly, if the language occurring in two different corpora is by definition not random, then to demonstrate that it is not random is meaningless – albeit one uses impressively sophisticated mathematics for this demonstration (Kilgarriff 2009). For this reason, in the research reported in this paper, the mathematics used to compare frequencies will be that of normalised frequencies, as Kilgarriff (2010: 3) recommends, without the use of chi-square, log likelihood, or any other probability test.

6.1 Phrasal verbs in interpreted English - frequency

The first finding, on the overall frequency of phrasal verbs in interpreted English, was that phrasal verb tokens were less frequent in INT-A than in the BNC. In the 100 million word BNC – as mentioned above – there is one phrasal verb every 192 words (Gardner and Davies 2007: 347), whereas the 364 tokens in INT-A work out as one phrasal verb approximately every 236 words. This raises the question of why, given the potential usefulness of phrasal verbs to interpreters, the frequency should be lower in interpreted English rather than higher. The logical explanation is the size factor. The hugely greater size of the BNC relative to INT-A means that it represents a much wider range of contexts. For example, from the point of view of mode, the BNC includes dialogue, whereas INT-A is limited to monologue. From the point of view of field, INT-A contains mainly parliamentary subject matter, whereas the BNC samples many different fields. From the genre angle, the BNC also contains many examples of texts exemplifying different literary genres, as well as journalism, both of which are absent from INT-A. In particular, it seems likely that the literature and journalism together with the dialogue present in the BNC will present narrative and interactional contexts of use of a large number of phrasal verb lemmas, whereas those same contexts of use are absent from the discourse sampled in INT-A.

A rate of one phrasal verb every 236 words is quite frequent, nevertheless. Assuming that interpreters towards English speak at 132–136 words per minute (Bendazzoli 2010: 149, 152), and taking the average of 134, it means interpreters were using on average one phrasal verb token approximately every 106 seconds; looked at another way, this means 136 phrasal verb tokens over four hours of interpreting. This is certainly a frequency high enough to confirm Hypothesis 1 (in section 3.4 above), that there would be a reasonably high occurrence of phrasal verbs in English produced by English native-speaker and nativelike simultaneous interpreters.

In terms of types, INT-A contained 131 phrasal verb lemmas. Table 1 shows the most frequent twenty-two, together with raw and normalised frequency (per million words), and, for the BNC, the frequency per million and the rank order of frequency. Additionally, the Table shows the ratio of keyness, obtained by dividing the normalised frequency of INT-A by the normalised frequency of the BNC. The keyness procedure is recommended by Kilgarriff (2010) as the most effective mathematical way to uncover differences between corpora.

Rank INT-A	Lemma	Rf INT-A	f/million INT-A	f/million BNC	Ratio Keyness	Rank Keyness	Rank BNC
1	carry out	28	326	108	3	11	2
2	move on	21	244	14	17.4	2	55
3	come up	17	198	55.2	3.6	10	10
3	draw up	17	198	25	7.9	6	0
5	set up	16	186	104	1.8	15	3

6	open up	15	174	21.4	8.1	5	0
7	take up	11	128	46	2.8	12	19
8	send out	10	116	13.5	8.6	4	0
9	bring in	9	105	25	4.2	9	37
10	point out	8	93	70	1.3	17	8
11	bring on	7	81	3.9	20.8	1	0
11	come back	7	81	80	1.0	20	6
13	sort out	6	70	27.8	2.5	13	0
13	pick up	6	70	90	0.8	22	4
15	come in	5	58	48	1.2	18	15
15	make up	5	58	55	1.1	19	11
15	put in	5	58	8	7.3	7	78
18	bring about	4	47	21	2.2	14	44
18	come out	4	47	50	0.9	21	13
18	end up	4	47	33	1.4	16	0
18	move in	4	47	8	5.9	8	79
18	send in	4	47	3.8	12.4	3	0

Table 1: The top 22 phrasal verbs in INT-A

As the rightmost column of Table 1 shows, fifteen of the most frequent phrasal verbs in INT-A also occurred in the Top 100 in the BNC as reported by Gardner and Davies (2007: 358). But Gardner and Davies' "top 100" is anomalous – it includes only phrasal verbs that contain one of the top twenty lexical verbs making up phrasal verbs, as contained in Gardner and Davies' Table 5 (Gardner and Davies 2007: 350). A search of the BYU-BNC shows that some phrasal verbs with lexical components outside the "lexical top twenty" are more frequent than the hundredth ranking phrasal verb in Gardner and Davies (2007), which has a frequency per million of 4.23 (see Gardner and Davies 2007: 359). Such verbs should therefore be considered as particularly frequent alongside the "top hundred". The f/million/BNC column of Table 1 shows that there are five such phrasal verbs in INT-A – *draw up*, *open up*, *send out*, *sort out* and *end up*. This means that a total of twenty out of twenty-two of the most frequent phrasal verbs in INT-A were among the most frequent in the BNC too. This in turn suggests that the phrasal verbs most often used in interpreted language are mostly used frequently in the language as a whole, although the differences are perhaps more interesting than the similarities. These differences are shown mathematically in Table 1 in the ratio of keyness column. The higher the ratio, the more characteristic the phrasal verb of interpreted English as represented in INT-A. According to this measure, the ten most characteristic verbs, in descending order, are *bring on*, *move on*, *send in*, *send out*, *open up*, *draw up*, *move in*, *bring in*, and *come up* (these verbs will be further discussed in sections 6.5 and 6.6). Examination of the contexts of some of the verbs that are ranked the highest in keyness in INT-A can help to define what marks out interpreted English, and to provide clues about formulae that are proven to be of use to interpreters. I will return to this topic in section 6.5.

6.2 Phrasal verbs in non-native interpreters: frequency

In INT_B, there were 59 phrasal verb tokens, representing one phrasal verb token every 409 words,

which means that the non-native interpreters used phrasal verbs as a category much less often than the interpreters of INT-A, and much less often than in the BNC, which confirms Hypothesis 2, and this in turn confirms the theory that processing difficulties make the choice of phrasal verbs by non-native interpreters less likely. It seems that, in the pressure of the SI situation the non-native interpreters had no time to search for phrasal verb items – items that, although probably known in theory, were hard to access in memory, probably due to their having being encountered only occasionally.

As far as individual phrasal verbs go, only 12 lemmas were used more than once. There was a considerable difference in the phrasal verb lemmas that were most frequent in INT-A and INT-B, as can be seen from Table 2, which shows the top 25 verbs in INT-A together with the rank orders, raw and normalised frequencies (per million) of the same verbs in INT-B, and the rank and ratio of keyness of the verbs in INT-B relative to the BNC. Only verbs with a ratio of keyness over 1.0 are ranked, as only these are statistically characteristic of INT-B relative to the language as a whole. All phrasal verbs in INT-B with a frequency of >1 are included in the Table.

LEMMA	Rank INT-A	Rank INT-B	Rf INT-B	F per million INT-B	Rank Keyness	Ratio Keyness
carry out	1	1	12	500	3	4.6
move on	2	7	2	83	2	5.6
come up	3	0	0	0	-	0
draw up	3	0	0	0	-	0
set up	5	4	4	167	6	1.6
open up	6	0	0	0	-	0
take up	7	0	0	0	-	0

send out	8	0	0	0	-	0.1
bring in	9	0	0	0	-	0
point out	10	13	1	42	-	0.6
bring on	11	7	0	0	-	0.2
come back	11	7	2	83	-	1.0
sort out	13	0	0	0	-	0
pick up	13	13	1	42	-	0.5
come in	15	13	1	42	-	0.9
make up	15	3	5	208	4	3.7
put in	15	0	0	0	-	0.9
bring about	18	0	0	0	-	0
come out	18	0	0	0	-	0
end up	18	0	0	0	-	0
move in	18	0	0	0	-	0.1
send in	18	0	0	0	-	0.2
go on	23	2	7	292	5	2.0
go back	23	5	3	125	7	1.5
sum up	23	5	3	125	1	10.2

Table 2: Comparison of phrasal verbs in INT-A and INT-B

Of course, it is easy to object to the finding that there was a lower frequency of phrasal verbs in INT-B than in INT-A on the grounds that the frequency differences are explicable in terms of the subject matter translated in INT-A being inherently more suitable for translation by phrasal verbs than the subject matter translated in INT-B. This can be tested through examination of one case, the material in DIRSI that is collected from two conferences on cystic fibrosis. Part of this material was translated into English by a NS interpreter, and part by NNS interpreters. Over 8436 words of text, the NS used 21 phrasal verb tokens, making one phrasal verb every 401 words on average, and covering 19 lemmas. The NNS interpreters used 22 phrasal verb tokens in 14.435 words, which makes one phrasal verb every 656 words, and covered 9 lemmas. At one every 401 words, the NS interpreter uses phrasal verbs less frequently than they are used in INT-A as a whole (see the first paragraph of section 6.1), so this would seem to confirm that discussions about cystic fibrosis do seem to require phrasal verb translations relatively seldom. On the other hand, this makes no difference to the relative frequency of phrasal verb use in the two corpora – frequency remains much higher in INT-A than it is in INT-B. So while it remains the case that the conference topic could have an effect, there is nevertheless a difference between the corpora, in terms of lower frequencies of phrasal verbs in INT-B as compared to INT-A . It is risky to restrict explanatory factors to one. So, given the evidence presented in section 3.2, the greater processing difficulties phrasal verbs present to non-native interpreters would seem to be a good candidate for an additional explanation for the lower frequency of phrasal verbs in INT-B.

6.3 Metafunctional distribution

To uncover information relating to Hypotheses 3 and 4, about phrasal verbs used for textual and interpersonal metafunctions, each phrasal verb concordance line was classified according to the metafunction that it predominantly expressed. The distribution of phrasal verb occurrences across the metafunctions in the two corpora is shown in Table 3.

Corpus	INT-A (n)	INT-B (n)	INT-A (%)	INT-B (%)
Metafunction				
Ideational	296	43	81	73
Interpersonal	18	11	5	19
Textual	50	5	14	8

Table 3: distribution of phrasal verbs across metafunctions

Table 3 shows that there was a substantial representation of all three metafunctions in the INT-A corpus. In the INT-B corpus, all three metafunctions were represented, but the proportion of phrasal verb tokens used for textual metafunction was lower than in INT-A. The data therefore confirmed Hypothesis 3, thus in turn confirming that native and nativelike interpreters use phrasal verbs as a resource to textualise their interpretations. Conversely, the data in terms of the much lower frequency of phrasal verbs used with textual metafunction in INT-B suggests that in tending not to use phrasal verbs when constructing the ongoing sense of their interpretations, non-native interpreters are neglecting a valuable resource.

For phrasal verbs used with interpersonal metafunction, the situation was reversed. The non-native interpreters used phrasal verbs interpersonally more often than the native speaker interpreters, so that Hypothesis 4 was not confirmed. It cannot entirely be discounted that the differing proportions of different types of message being translated (for example, allocation of seating, transitions of speaker, management of latecomers – which are more likely to involve mediation) may have influenced the disparity in frequency of interpersonal phrasal verbs between the two corpora. However this cannot reliably be checked, since there was no information of this type in 2249,

which formed a substantial component of INT-A. An alternative explanation is that the non-native interpreters have substantial experience of phrasal verbs used interpersonally, as a result of having passed through courses of English language conducted using communicative language teaching methods. Therefore the assumption that phrasal verbs would be hard to access for non-native interpreters during SI does not apply in the case of interpersonal contexts.

In sections 6.4 to 6.7, there is a detailed examination of examples of phrasal verbs used for the three metafunctions, which reveals the prominent role carried out by phrasal verbs in the work of the interpreter. The examples also demonstrate the functions that the phrasal verbs are being used for.

6.4 Mediation using interpersonal phrasal verbs

When used with interpersonal metafunction, phrasal verbs were used to expedite the interpreter's role of mediator, in helping to make a third party do something or feel something. The following examples from INT-A show how phrasal verbs express the process of mediation:

(5) I will *give* the floor *back* to you.

(6) I could *hand out* some of the tables with the relevant information

(7) I'd like to thank you for … the fact that you've *stuck* it *out* this evening.

(8) could you please *come in* and close the door.

As Table 3 shows, phrasal verbs with interactional metafunction were relatively well represented in INT-B, with examples like the following:

(9) please *hand back* the headset and receivers for the interpretation to the desk at the entrance

(10) just five minutes while *filling in* the questionnaire and *pick up* the devices the headsets

All of these examples show how phrasal verbs are used by interpreters in their integral role of mediator, to expedite event proceedings. It is interesting to note that phrasal verbs are used in this role in both types of event, parliamentary debates (5, 6, 7) and conferences (8), represented in INT-A, as well as in the conferences represented in INT-B (9 and 10).

6.5 Textualising phrasal verbs in interpreted language

Phrasal verbs used with textual metafunction are also indexical to the interpreter's role, being involved in the interpreting process in the sense of Setton's account of online meaning assembly, which sees simultaneous interpreters as involved in a continuous process of simulating and recreating context:

> relevance (coherence) is sought as a matter of routine. Context is seen as a nested set of assumptions: a background about the world and the situation, then assumptions based on previous discourse, on the previous utterance, and finally those taking shape from the utterance-initial cues about the illocution or modality of the ongoing argument […] (Setton 1998/2002: 193).

The way that phrasal verbs convey the process of creation and articulation of textual meaning in SI is shown in examples (11) to (19) from INT-A. Here is an account of the rhetorical function of each example.

In examples (11) and (12) , the interpreter is indicating to herself and to the hearers that between propositions a relation of evidence should be perceived; in (13), (14) and (15) there is a direction to perceive a relation of presentation; in (16) there is the perception and relaying of a relation of recap; in (17) with both the initial choice *focus down* and the reformulation *narrow down* there is the evocation of a relation of elaboration involving a perception of movement from the general to the specific; and (18) and (19) are relations of resumption **[1]**.

(11) The + data can all be documented, to *back up* what I said.

(12) this analysis has in fact been *borne out* by a number of different contributions

(13) That *brings* us *on* to the report, by Mrs Lucas

(14) we're *kicking off* on the basis of an assessment report

(15) We *move on* to the report by Mr Sakalas.

(16) And that *brings* me *back* to what I was saying at the very beginning

(17) this allowed us to *focus down* to *narrow down* the field to

(18) Now we'll *go on* with the er debate

(19) Well, I would like to er *pick up* on the point on the credit rating agencies.

Bring on is the highest ranking phrasal verb by keyness in INT-A, while *move on* ranks second in keyness in INT-A (see Table 1 above), and first in INT-B (see Table 2). Six of the seven occurrences of *bring on* in INT-A express the textual metafunction, as do fifteen of the twenty-one occurrences of *move on* – so these would seem to be two instances of relative prominence of a phrasal verb in INT-A which is explicable in terms of aspects of the process of interpreting, in this case textualisation.

Phrasal verbs used for the textual metafunction were much less common in INT-B. However, the two most key phrasal verbs in INT-B (see Table 2) were used with textual metafunction. These were *move on*, signalling the presentation relation, and *sum up*, used to indicate a relation of summary. It is thus clear that textualisation of output was taking place explicitly in INT-B, as one would expect. Given this, it could be that different types of conference situation provide the explanation for the lower frequency of textualising phrasal verbs in INT-B; in INT-A, there is a preponderance of parliamentary debates, whereas in INT-B there is a preponderance of conferences with the function of updating members of a specific professional community. But given that message construction with attendant textualisation should be taking place in both contexts, it is equally plausible to explain the low frequency of phrasal verbs with textual metafunction in INT-B through a lack of experience on the part of the non-native interpreters of textualising in English, or simply through the pressure of the SI situation, both of which plausibly result in the signalling of textual structure through structures transferable from the mother tongue, rather than through phrasal verbs.

6.6 Ideational phrasal verbs in interpreted language

The ideational metafunction, by virtue of its application to content, can be stated to cover interpreted product. Phrasal verbs used for the ideational function in INT-A cover a wide range of types of meaning. As indicated above, some phrasal verbs used with ideational meaning are likely to recur whatever the subject matter being translated, such as certain mental process verbs, while others will be linked with a more restricted, specialised context. The discussion is organised according to verb types – mental processes, material processes, and relational processes. The degree of productive value for interpreters of certain phrasal verbs, and of the multiword expressions of which they form part, as revealed by the corpora, is discussed within these sections. Of course verbs cannot intrinsically be placed in a single one of the three categories of mental, material and relational processes. I classified them as belonging to one of the three categories on the basis of the type of meaning that they predominantly expressed in each single instance of use in context actually observed in the KWIC concordance.

Starting with mental processes, some of the more frequent phrasal verbs in INT-A form components of collocations or slot and filler patterns that extend the phrasal verb into longer multiword expressions which are likely to be formulaic sequences particularly worth learning for non-native interpreters. There is for example the phrasal-prepositional verb *come up with*, which occurred as many as fifteen times, thus contributing to this verb's position as number ten for keyness in INT-A (see Table 1). *Come up with* collocated relatively frequently (seven times) with *proposal* (example 20).

(20) we asked the Commission to *come up with a proposal* to harmonise the guarantee.

At a more general level of semantic prosody, a slot and filler pattern can be observed, which is *come up with* + **noun phrase describing something viewed positively in the context.** Thirteen of the occurrences of *come up with* fall into this category. A further specification for this particular formulaic sequence is that the positive semantic prosody is reflected in ten cases out of fifteen by collocates that are positive adjectives (*excellent, good, speedy, flexible*) – see (21).

(21) we've been able to *come up with a very pragmatic and flexible* proposal.

Although the mental nature of the process represented by *come up with* would suggest it would occur in a wide range of situations, in INT-A it only occurred in parliamentary debates. This is almost certainly a reflection of the small proportion of INT-A (about 10 per cent) represented by conference interpreting. It is true that *come up with*, through its association with solutions to problems, is closely associated with parliamentary process; but it is also associated with idiom, so it should occur fairly frequently in conferences, given that they are often held for the purpose of exchanging information about innovations. This unexpected lack of instances reflects the need for bigger samples of conference interpreting, to produce more representative results.

Moving on to material processes, *carry out* was also observed to combine in INT-A with collocates in formulaic sequences, notably with *analysis* (three times), *checks* (three times), *a study/studies* (three times), *actions* (twice) and **noun phrase modifier** + *activities* (twice) – for example*, carrying out the reception activities*). This was one of the few verbs to occur really frequently in INT-B (12 occurrences), where it was used with *study/studies* (twice), *demonstrations* and *research*. This verb was used in both the parliamentary debates and the conference speeches, so it would seem to be particularly productive for interpreters, in the sense of having a wide application.

The material process verb *set up* was also fairly frequent in both INT-A and INT-B. In INT-A there are few indications of extended multiword expressions, with *mechanism(s)* being the only collocate of *set up* that recurs (twice). There were no repeated collocations in INT-B, though the contexts of use were standard (*access*, *laboratory*, *model*), so that one can infer that in a larger corpus of non-native interpreting, some recurring collocations would have emerged.

The material process verbs *draw up* and *open up*, as shown by their contexts viewed in the KWIC concordances, are productive in the context of interpreting in the sense that they refer to content that interpreters in international institutions are likely to repeatedly experience and then produce. The noun collocates of *draw up* in INT-A are *law* (twice), *regulation*, *list*, *plan* and *agreement*. All of the occurrences were in parliamentary debates, and closely reflect the role of parliament. It is therefore unsurprising that *draw up* did not occur in INT-B, which is almost entirely composed of medical conference interpretations. The association of *draw up* with parliamentary procedure explains why it ranked as high as sixth in terms of keyness in INT-A as compared with the BNC (see Table 1 above). The noun collocates of *open up* are mostly (nine times) the word *market* or synonyms of *market*, reflecting the current preoccupation of the European parliament with market-led economics, which in turn explains the number five ranking of this item in the INT-A keyness scale (see Table 1). *Open up* was absent from INT-B, which is unsurprising given that almost all of the subject matter is concerned with social care and health matters rather than economics. But it is surely uncontroversial to claim that, for as long as the hegemony of market economics persists*, open up + market(s)* will be a productive multiword pattern for non-native interpreters to learn to automatically produce.

Relational processes are quite frequently conveyed in INT-A through phrasal verbs, for example *end up* (four times), *bring about* (four times), *make up* – meaning *constitute* – three times, and *bring in* (nine times). One formulaic sequence was discernible in the case of *bring in*, which is its collocation with *new* (three times), and, more generally, there was an association with the semantic field of innovation (eight times), involving adjectives like *extraordinary* and nouns like *improvement* and *advances*. *Make up* (=constitute) also occurred three times in INT-B, with standard contexts of use (e.g. *an audience made up of friends*), but none of the collocations was repeated.

So far the ideational phrasal verbs reported on here are shown by their presence in the BNC top 100 to be reasonably frequent in the language as a whole. But (as Table 1 shows, and as mentioned in section 6.1) there is also the question of the keyness of some items, in other words of their conspicuously high normalised frequencies in INT-A as compared with the BNC. Among these are *bring on*, *move on*, *come up*, *draw up* and *open up,* whose keyness has already been discussed. For the remaining members of the top ten of keyness in INT-A (see Table 1), *send in*, *send out, put in*, *move in* and *bring in*, it would be tempting to claim that their higher frequency in INT-A makes them typical of interpreted English. However, there is nothing in the contexts revealed by the concordances to support this, and all of the occurrences are from parliamentary debates. So it is more realistic to attribute the higher frequencies of these phrasal verbs to the fact that the contexts where they were the appropriate translations happened to have occurred in parliamentary business in the debates concerned.

It is interesting to note that there is a reasonably wide variety of ideational phrasal verbs in INT-B, with 16 different types represented. This seems to suggest that the non-native interpreters are not inherently shy about using phrasal verbs for translated subject matter, albeit the figures show that they use them less frequently than native speaker and native-like interpreters. It is also in contrast with the lack of variety (only two types, *sum up* and *move*

on) of phrasal verbs used in INT-B for the textual metafunction. Together with the low number of tokens, this suggests a precise reason why so few phrasal verb types and tokens with textual metafunction are used by non-native interpreters. This reason is that textualisation is the aspect of interpreting over which interpreters have most control. By contrast, for ideational function (i.e. content) the context set by the original speaker is probably steering the interpreter into the phrasal verbs associated with it. In this respect, with ideational function, non-native interpreters appear to be behaving like native and native-like interpreters, in using phrasal verbs as a processing resource, rather than experiencing them as a processing difficulty (though, the frequency figures suggest, with non-native interpreters this happens much less often). But in the textualisation process, with the creation of the textual framework the responsibility of the interpreter, the non-native interpreters are probably automatically falling back on textualising structures analogous to L1, and are thus avoiding phrasal verbs in the process. This is speculation, of course, but it would be interesting if some research could be designed to test this idea.

6.7 Idiomatic/figurative and aspectual phrasal verbs in INT-A and INT-B

The data on literal, idiomatic/figurative and aspectual phrasal verbs in INT-A and INT-B is summarised in Table 4. The verbs listed by name in the Table are the most frequent in each category. An asterisk indicates that the verb is neither present in Gardner and Davies' BNC top 100 nor has a BYU-BNC frequency high enough to be included in it, in other words a frequency of 4.23 per million, that of Gardner and Davies' hundredth ranked item. The data confirmed Hypothesis 5, which predicted that non-native interpreters working towards English as their B language would use idiomatic and figurative phrasal verbs less than do native and native-like interpreters. In INT-A, there were 2200 figurative/idiomatic phrasal verb tokens per million words, while in INT-B there were only 1200 tokens per 1000 words. These data reinforce the theory that processing difficulties make idiomatic and figurative phrasal verbs difficult to access for non-native interpreters working towards B.

Table 4 shows that, as might be expected in the circumstances, the number of different lemmas used with figurative/idiomatic meaning was much higher in INT-A at 75, while in INT-B the total was 8, which is surprisingly low even considering that the number of words in INT-B was less than a third of the number of words in INT-A. Indeed, in INT-B 12 of the 29 figurative/idiomatic tokens are accounted for by a single lemma, *carry out*, and all but one figurative/idiomatic lemmas are inherently frequent, as shown by their having a frequency per million in the BNC above the threshold of 4.23. This suggests that access to individual figurative/idiomatic phrasal verbs during B-direction interpreting is likely only if frequent experience has brought them relatively near to the surface of memory. Conversely, the finding that figurative meanings of phrasal verbs less frequent than the 4.23 per million threshold are hardly used at all by the B-direction interpreters (only a single occurrence of *spring up* – see Table 4) has a plausible explanation. It is likely that, given the time lag shown in the literature for NNS access to non-literal items, less frequent phrasal verbs (even though they may be known) take too long for the non-native to access in the SI situation.

	Figurative / Idiomatic	**Aspectual**	**Literal**
INT-A - raw F tokens	192	55	117
INT-B - raw F tokens	29	2	28
INT-A - tokens/million words	2200	600	1400
INT-B - tokens/million words	1200	100	1200
INT-A - lemmas (n)	75	25	41
INT-B- lemmas (n)	8	2	
INT-A - lemmas ranked 1–8 (*Rf*, f/million words)	carry out (*30,* 349.3), come up (*17,* 197.9), set up (*16,* 186.3), take up (*9,* 104.8), point out (*8,* 93.2), draw up (*6,* 69.9), make up *(5,* 58.2), pick up *(5,* 58.2)	open up (*15,* 174.6), sort out (*6,* 69.9), end up (*4,* 46.6), slow down *(3,* 34.9), sum up (*3,* 34.9), find out (*2,* 23.3), help out (*2,* 23.3), start out (*2,* 23.3), weigh up* (*2,* 23.3)	move on (*20,* 232.9), send out (*10,* 116.4), bring in (*9,* 104.8), bring on (*6,* 69.9), come back (*6,* 69.9), move in (*5,* 58.2), come in (*4,* 46.6), go down (*4,* 46.6), move around (*4*, 46.6), send in *(4,* 46.6)
INT-B - lemmas ranked 1–8(*Rf*, f/million words)	carry out (*12,* 497.5), make up (*5,* 207.3), set up (*4,* 165.8), sum up (3, 124.4), go on (2, 82.9), bring up (*1*, 41.5), point out (*1*, 41.5), spring up* (*1*, 41.5)	find out (*1,* 41.5), fill out*, (*1,* 41.5)	go on (*4,* 165.8), go back *(3,* 124.4), come back (*2,* 82.9), fill in (*2,* 82.9), get back (*2,* 82.9), keep on (*2,* 82.9), miss out (*2,* 82.9) , move on (*2,* 82.9)

Key - * phrasal verb with a frequency in the BNC of <4.23 per million words

Table 4: Idiomatic/figurative, aspectual and literal phrasal verbs

In contrast, inherently less frequent non-literal phrasal verbs did occur in INT-A, with the presence (mostly as one-off occurrences) both of non-compositional lemmas like *beef up, crop up, kick in* and *stump up*, and of lemmas used figuratively, such as *chime in, float around, iron out, thrash out* and *whittle down*, to name some examples of verbs found on the BYU-BNC to have frequencies lower than the 4.23 per million threshold. This occurrence of not particularly frequent idiomatic and figurative items in the INT-A corpus is another way in which the corpus shows that use of phrasal verbs in interpreted language resembles the use of phrasal verbs in the language as a whole, and is consonant with the theory that NS and native-like interpreters, like English speakers in general, derive processing advantages from the use of such idiomatic/figurative phrases.

The figures for aspectual phrasal verbs follow a similar pattern, with frequencies of aspectual tokens per million words much higher in INT A (600) than in INT B (100). Once again, the variety of lemmas was disproportionately lower in INT-B, with only two lemmas used with aspectual meaning in INT-B as against twenty-five in INT-A. The suggestion here once again is that phrasal verbs with aspectual meaning do not come readily to the mind of the B-direction interpreter engaged in SI; and that a phrasal resource that is used to advantage by native and native-like interpreters, seems plausibly to constitute a difficulty, to be circumvented, for the non-native interpreter.

7. Conclusions

In this article I have set out to make the case for the teaching of phrasal verbs to trainee interpreters. The case rests on six findings.

1. The observation of reasonably frequent use of phrasal verbs by native and native-like interpreters in INT-A.
2. The observed use of phrasal verbs by the native and native-like interpreters in INT-A for the textualisation that is a crucial aspect of the SI process.
3. The observed use of phrasal verbs for interpreter-mediated interpersonal functions.
4. The observed use of phrasal verbs for mental processes which are likely to recur in situations where interpreting takes place.

Findings 1–4 are intended to show that phrasal verbs are part of the linguistic fabric of professional SI. Findings 5 and 6 concern non-native interpreters.

5. There was confirmation of the hypothesis that processing difficulties in SI would lead to a lower frequency of phrasal verb use by the non-native interpreters of INT-B.
6. There were very much lower frequencies in INT-B of phrasal verbs with textual metafunction, which demonstrate a linguistic lack, and a consequent need for learning and teaching (though there appears to be less need for instruction on phrasal verbs with interpersonal metafunction).

It is conceded that the small size of the INT-B corpus, and the restriction of the corpus to only part of three conferences with just a small section of parliamentary interpreting, must make finding number 5, that there is a relatively low frequency of phrasal verbs by non-native speakers, a tentative one. The inclusion of different subject matter, forming a larger sample, might have revealed more occasions when opportunities were taken up by interpreters for the choice of phrasal verbs in translations. Finding number 1, that there was a reasonably high frequency of phrasal verbs in INT-A, should also be received with a degree of caution, due to the low proportion of conference interpreting and the preponderance in the data of parliamentary debates. Overall, however, the results found in both INT-A and in INT-B are in accord with the theories that propose that native speakers find processing advantages in phrasal verbs while non-natives find processing difficulties, and for that reason it seems likely that replication of the research with a larger and more varied non-native interpreter corpus would result in similar findings.

The lower frequencies of phrasal verbs in the non-native interpreters corpus suggests that there is room for improvement in phrasal verb knowledge if non-natives are to more nearly approach native or native-like standards in terms of readiness of automatised formulaic sequences involving phrasal verbs during meaning assembly when working from A to B. Similarly, a need for improved knowledge is indicated by the low frequency of non-literal phrasal verbs among the non-native interpreters, which if translated from production to reception, would mean the risk of non-comprehension when working from B to A. The literature suggests that, partly because of the large number of non-literal phrasal verbs, acquisition cannot be left to simple experience, even when residence in an English speaking country is involved (Siyanova and Schmitt 2007), and this in turn suggests that for improvement in knowledge of phrasal verbs to take place, there must be some form of designed instruction, preferably before trainees begin interpreting work, and preferably involving attentional processes (Schmidt 1990). This form of instruction cannot take place in interpreting practice classes, precisely because attention there is perforce directed away from linguistic form; so it should take place in adjunct language support courses. Material for such courses has been uncovered in the current study, where analysis of the KWIC concordances revealed a number of longer multiword formulae (phrasal verbs +…) useful for interpreted language. Some of the concordances have already been converted into learning materials and used with trainees at the Department of Interpreting and Translation (University of Bologna at Forlì). The hope is that such work will continue, and that our understanding of the role of formulaic language in the process and product of interpreted English can be extended by the development of corpus research methods. This should include more representation in corpora of conference interpreting, to take its place alongside existing corpora of interpretation of parliamentary debates. Work also needs to be done on developing larger corpora of non-native interpreters. The small size of the non-native interpreted language corpus used in this research has meant that the findings presented here must be qualified as merely tentative. But this is no reason to disregard them, for progress in corpus research on interpreted language has to start somewhere, and in any case, it is unrealistic to expect this progress, particularly in a relatively unexplored field, to be anything other than gradual and incremental.

References

Aston, Guy (1997) “Small and Large Corpora in Language Learning”, paper presented at the PALC conference, Lodz, URL: **http: //www.sslmit.unibo.it/~guy/wudj1.htm** (accessed 16 January 2015).

---- (2017) “Acquiring the Language of Interpreters: A Corpus-based Approach,” in *Making Way in Corpus-based Interpreting Studies*, Mariachiara Russo, Claudio Bendazzoli and Bart Defrancq (eds), Singapore, Springer.

Bartłomiejczyk, Magdalena (2004) “Simultaneous Interpreting A-B vs. B-A from the Interpreters’ Standpoint” in *Claims, Changes and Challenges in Translation Studies - Selected contributions from the EST Congress, Copenhagen 2001*, Gyde Hansen, Kirsten Malmkjaer and Daniel Gile (eds), Amsterdam, John Benjamins: 239–250.

Bendazzoli, Claudio (2010) *Corpora e Interpretazione Simultanea,* Bologna, Asterisco Edizioni.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999) *Longman Grammar of Spoken and Written English*, Harlow, Longman.

Celce-Murcia, Marianne, and Diane Larsen-Freeman (1999) *The Grammar Book: An ESL/EFL Teacher’s Course* (2nd ed.), Boston, Heinle & Heinle.

Cieślicka, Anna (2006) “Literal Salience in On-line Processing of Idiomatic Expressions by Second Language Learners”, *Second Language Research* 22, no. 2: 115–44.

Cresswell, Andy (2013) “Creating KWIC-searchable Corpora Through Tagging Argumentative Texts with Logical Relations Using Rhetorical Structure Theory: Relation Definitions and Discourse Unit Division Protocols,” Bologna, AMS Acta - Research repository of the University of Bologna, URL: **http: //amsacta.unibo.it/3654** (accessed 20 June 2016).

Dagut, Menachem and Batia Laufer (1985) “Avoidance of Phrasal Verbs: A Case for Contrastive Analysis”, *Studies in Second Language Acquisition* 7, no.1: 73–79.

Darwin, Clayton and Loretta S. Gray (1999) “Going After the Phrasal Verb: An Alternative Approach to Classification”, *TESOL Quarterly* 33, no. 1: 65–83.

Davies, Mark (2004) *BYU-BNC* (Based on the British National Corpus from Oxford University Press), URL: **http://corpus.byu.edu/bnc/** (accessed 12 May 2017).

Denissenko, Jurij (1989) “Communicative and Interpretative Linguistics” in *The Theoretical and Practical Aspects of Teaching Conference Interpretation*, Laura Gran and John M. Dodds (eds), Udine, Campanotto Editore: 155–158.

Flowerdew, Lynne (2004) “The Argument for Using English Specialized Corpora to Understand Academic and Professional Language” in *Discourse in the Professions*, Ulla Connor and Thomas Upton (eds), Amsterdam, John Benjamins: 11–33.

Gardner, Dee and Mark Davies (2007) “Pointing out Frequent Phrasal Verbs: A Corpus-based Analysis”, *Tesol Quarterly* 41, no.2: 339–359.

Grant, Lynn and Laurie Bauer (2004) “Criteria for Re-defining Idioms: Are We Barking up the Wrong Tree?”, *Applied Linguistics* 25, no.1: 38–61.

Halliday, Michael Alexander Kirkwood, and Christian Matthias Ingemar Martin Matthiessen (2014) *Halliday’s Introduction to Functional Grammar (4th edition)*, London, Routledge.

Herbert, Jean (1952) *Manual del Intérprete. Del Entrenamiento para Llegar a Intérprete de Conferencias*, Geneva, Georg.

Hood, Sue, and Gail Forey (1999) “Research to Pedagogy in EAP Literature Reviews”, paper presented at TESOL convention, March 1999, New York.

Hulstijn, Jan, and Elaine Marchena (1989) “Avoidance: Grammatical or Semantic Causes?” *Studies in Second Language* Acquisition 11, no. 3: 241–255.

Hyland, Ken (2005) *Metadiscourse: Exploring Interaction in Writing*, London, Continuum.

Jenkins, Jennifer (2001) “Euro-English Accents”, *English Today* 68, no.17: 16–19.

Kilgarriff, Adam (2001) “Comparing Corpora”, *International Journal of Corpus Linguistics* 6, no. 1: 97–133.

---- (2005) “Language is Never, Ever, Ever Random”, *Corpus Linguistics and Linguistic Theory* 1, no. 2: 263–275.

---- (2009), “Simple Maths for Key Words” in Michaela Mahlberg, Victorina Gonzàlez-Dìaz and Catherine Smith (eds), *Proceedings of the Corpus Linguistics Conference CL2009*, Liverpool, University of Liverpool, URL: **http://ucrel.lancs.ac.uk/publications/cl2009/** (accessed 12 May 2017).

---- (2010) “Comparable Corpora Within and Across Languages, Word Frequency Lists and the Kelly Project”, invited talk at the LREC workshop on Building and Using Comparable Corpora, Malta, May 2010, URL: **https://www.sketchengine.co.uk/wp-content/uploads/Comparable_corpora_within_2010.pdf** (accessed 12 May 2017).

Lakoff, George (1986) “The Meanings of Literal”, *Metaphor and Symbolic Activity* 1, no. 4: 291–6.

Laufer, Batia (1997) “What’s in a Word that Makes it Hard or Easy: Some Intralexical Factors that Affect the Learning of Words” in *Vocabulary: Description, Acquisition and Pedagogy*, Norbert Schmitt and Michael McCarthy (eds), Cambridge, Cambridge University Press: 140–155.

Laufer, Batia, and Stig Eliasson (1993) “What Causes Avoidance in L2 Learning: L1-L2 Difference, L1–L2 Similarity or L2 Complexity?” *Studies in Second Language Acquisition* 15, no. 1: 35–48.

Liao, Yan D. and Yoshinori Fukuya J. (2004) “Avoidance of Phrasal Verbs: The Case of Chinese Learners of English”, *Language Learning* 54, no. 2: 193–226.

Lyons, John (1995) *Linguistic Semantics. An Introduction*, Cambridge, Cambridge University Press.

Martin, James R. (1992) *English Text: System and Structure*, Amsterdam, John Benjamins.

McArthur, Tom (2003) “World English, Euro-English, Nordic English?”, *English Today* 19, no. 1: 54–58.

Modiano, Marko (2003) “Euro-English: a Swedish perspective”, *English Today* 19, no. 2: 35–41.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli, and Mariachiara Russo (2005) “Studying Directionality in Simultaneous Interpreting Through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus)”, *Meta*

50, no. 4, URL: **https: //www.erudit.org/revue/meta/2005/v50/n4/019850ar.pdf** (accessed 6 November 2016).

Moon, Rosamund (1997) "Vocabulary Connections: Multi-word Items in English" in *Vocabulary: Description, Acquisition and Pedagogy,* Norbert Schmitt and Michael McCarthy (eds), Cambridge, Cambridge University Press: 140–155.

Pawley, Andrew, and Frances Hodgetts Syder (1983) "Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency" in *Language and Communication,* Jack C. Richards and Richard W. Schmidt (eds), London, Longman: 191–226.

---- (2000) "The One Clause at a Time Hypothesis" in *Perspectives on Fluency*, Heidi Riggenbach (ed.), Ann Arbor, University of Michigan Press: 163–199.

Rost, Michael (2011) *Teaching and Researching Listening*, Harlow, Longman.

Sandrelli, Annalisa, Claudio Bendazzoli, and Mariachiara Russo (2010) "European Parliament Interpreting Corpus (EPIC): Methodological issues and preliminary results on lexical patterns in SI", *International Journal of Translation* 22, no. 1/2: 165–203.

Schmidt, Richard W. (1990) "The Role of Consciousness in Second Language Learning", *Applied Linguistics*, 11, no. 2: 127–158.

Schneider, E.W. (2004) "How to Trace Structural Nativization: Particle Verbs in World Englishes", *World Englishes* 23, no. 2: 227–249.

Scott, Michael (2012/16) *Wordmith Tools*, *version 6/7*, Stroud, Lexical Analysis Software. URL: **www.lexically.net** (accessed 19 June 2016).

Seidlhofer, Barbara (2007) "Common Property: English as a Lingua Franca in Europe" in *International Handbook of English Language Teaching,* Jim Cummins and Chris Davison (eds), Springer, New York: 137–153.

Seleskovitch, Danica (1978) *Interpreting for International Conferences: Problems of Language and Communication*, Washington, D.C., Pen and Booth. Originally published 1968 as *L'Interprète dans les conferences internationals: problèmes de langue et de communication*, Paris, Lettres Modernes Minard.

Setton, Robin (1998/2002) "Meaning Assembly in Simultaneous Translation" in *The Interpreting Studies Reader*, Franz Pöchhacker and Miriam Shlesinger (eds), London, Routledge: 178–202.

Side, Richard (1990) "Phrasal Verbs: Sorting Them Out", *ELT Journal* 44, no. 2: 144–152.

Siyanova, Anna and Norbert Schmitt (2007) "Native and Nonnative Use of Multiword vs. One-word Verbs", *International Review of Applied Linguistics* 45: 119–139.

Siyanova-Chanturia Anna, and Ron Martinez (2015) "The Idiom Principle Revisited", *Applied Linguistics* 36, no. 5: 549–69.

Siyanova-Chanturia, Anna, Kathy Conklin, and Norbert Schmitt (2011) "Adding More Fuel to the Fire: An Eye-tracking Study of Idiom Processing by Native and Non-native Speakers", *Second Language Research* 27, no. 2: 251–72.

Trebits, Anna (2009) "The Most Frequent Phrasal Verbs in English Language EU Documents – A Corpus-based Analysis and its Implications", *System* 37, no. 3: 470–81.

Underwood, Geoffrey, Norbert Schmitt, and Adam Galpin (2004) "The Eyes Have It: An Eye-movement Study into the Processing of Formulaic Sequences" in *Formulaic Sequences*, Norbert Schmitt (ed), Amsterdam, John Benjamins: 153–72.

Wray, Alison (2000) "Formulaic Sequences in Second Language Teaching: Principle and Practice", *Applied Linguistics,* 21, no. 4: 463–89.

---- (2002) *Formulaic Language and the Lexicon,* Cambridge, Cambridge University Press.

Notes

[1] For a more detailed account of the pragmatics of these and other "coherence relations", see Cresswell (2013).

"Looking up phrasal verbs in small corpora of interpreting An attempt to draw out aspects of interpreted language", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.

Stable URL: http://www.intralinea.org/specials/article/2319

"*ANGLINTRAD*: Towards a purpose specific interpreting corpus", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2317

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy.

ANGLINTRAD: Towards a purpose specific interpreting corpus

By Michela Bertozzi (Università di Bologna, Italy)

Abstract & Keywords

English:

Corpus-based interpreting methods are effective for analyzing important phenomena that has been neglected in research (Shlesinger 1998), but little attention has been paid to their possible exploitation in interpreter education and to the benefits of corpus-derived insights for trainee interpreters (Bendazzoli 2010a). The aim of this paper is to describe how *Anglintrad*, a purpose-specific intermodal Italian-Spanish corpus, is being built and to suggest some preliminary exploitation criteria for interpreter (and translator) training and practice. This paper focuses on the presence of unmodified English loanwords in Italian political speeches (Marzocchi 2007) and their renditions in simultaneous interpreting and written translation into Spanish. The possibility of comparing the same phenomena (unmodified English loanwords) from two different perspectives (interpreting and translation) represents an unprecedented opportunity entailing possible didactic applications to enhance interpreter and translator training and practice.

Keywords: interpreting, anglicisms, corpora, loanwords, italian, spanish

1. Introduction

Over the last decades, corpus-based and corpus-driven interpreting studies (CIS) have significantly developed from the ground-breaking research on the first manual corpora of courtroom interpreting by Shlesinger (1989), the mid-late nineties with an increasing number of studies on interpreting corpora by Pöchhacker (1994), Kalina (1998), Setton (1997, 1999) and the years between the late nineties and the beginning of the third millennium, characterized by Shlesinger's plea (1998) to make research efforts into the compilation and use of electronic, machine-readable interpreting corpora:

> From the standpoint of interpreting research, the compilation of bilingual and parallel corpora is indeed overdue, given the potential to use large, machine-readable corpora to arrive at global inferences about the interpreted text in relation to other forms of oral discourse; and in relation to other forms of translation.(Shlesinger 1998: 2)

This paved the way for a new approach in interpreting research, where the methodology of Corpus Linguistics was applied to the creation and consultation of the first machine-readable interpreting corpora (Cencini and Aston 2002, Wallmach 2002, Bendazzoli *et al*. 2004, Timarova 2005, Shlesinger 2008). Over the last few years, researchers have been channeling their efforts towards open-access electronic corpora (House, Meyer and Schmidt 2012, Monti *et al*. 2005, Bendazzoli and Sandrelli 2005-2007, Sandrelli *et al*. 2010).

However, so far little attention has been paid to the possible exploitation of these corpora for interpreter training and the benefits of corpus-derived insights for didactic purposes. Providing interpreting (and translation) trainees with a user-friendly platform, for instance collecting data on unmodified English loanwords, would be beneficial from a didactic point of view for several reasons: first, raising awareness on the issue of unmodified English loanwords in Italian and how this phenomenon can be managed in interpreting and translation; second, providing students with a set of different strategies applied by professionals in a high-quality, homogenous and comparable setting can bring added value to interpreting and translation training sessions, which would entail the possibility to compare the trainee's renditions (or translations) with the professional ones. As a matter of fact, the speeches, interpretations and translations at European Parliament have already been used for teaching purposes as a source of didactic material and this corpus could be an extra tool to be used both by teachers and students in their training sessions; finally, a platform comparing the same phenomenon from the interpreter and translator's point of view could be exploited to make trainees expand their own perspective on the array of possible strategies that can (or cannot) be used both in interpreting and translation.

1.1 Objectives

The aim of the present study is to present the methodology and contents of *Anglintrad**[1]***, a purpose-specific intermodal (interpreting and translation) Italian-Spanish corpus, and to highlight some preliminary didactic implications for future interpreters and translators.

The idea of the *Anglintrad* corpus came from the practical need to shed light on a particularly challenging phenomenon in Italian-Spanish simultaneous interpreting, that is the frequent use of unmodified English loanwords**[2]** in Italian political speeches (Marzocchi 2007) and the different Spanish mechanisms of loanword integration (Tonin 2010); these phenomena have been widely studied in translation, but little attention has been paid to understanding how they can affect the interpreter's performance. Therefore, *Anglintrad* was specifically designed with a view to selecting a number of oral texts delivered within the same setting (the European Parliament plenary sittings) sharing a common characteristic (the presence of unmodified English loanwords in the original Italian speeches), then comparing them with the corresponding Spanish interpreted speeches and official translations. The fact that the corpus is intermodal (including both interpreted and translated texts) may lead to future comparative studies, as already suggested by Shlesinger's paper on the comparison between written and oral corpora:

> Ideally, the notion of comparable corpora in interpreting studies should be extended to cover setting up three separate collections of texts in the same language: interpreted texts, original oral discourses delivered in similar settings, and

> written translations of such texts. This would allow for the identification of patterns specific to interpreted texts (regardless of their source language) as pieces of oral discourse, in relation to comparable texts in the same language. It would also allow us to identify the patterns which single out interpreted texts as distinct oral translational products in a given language irrespective of their source languages, through comparisons with comparable written translational products. (Shlesinger 1998: 4)

In the light of the above, the ultimate goal of the *Anglintrad* project is a bilingual intermodal corpus to observe a particular phenomenon (the presence of unmodified English loanwords in Italian original speeches delivered in the European Parliament plenary sitting), the way it is managed by simultaneous interpreters into Spanish and by translators into the same target language not only within the same setting (the plenary sitting itself) but within the same original text that is studied from two different perspectives.

1.2 Corpus structure

The *Anglintrad* corpus is divided into two main sub-corpora: oral (1) and written (2) texts (see Figure 1). The former includes original Italian speeches delivered at the European Parliament plenary sitting in the year 2011 (1A) with the related interpreted Spanish versions (1B); the latter is made up of the official revised Spanish translations referred to the same original speeches (2A).

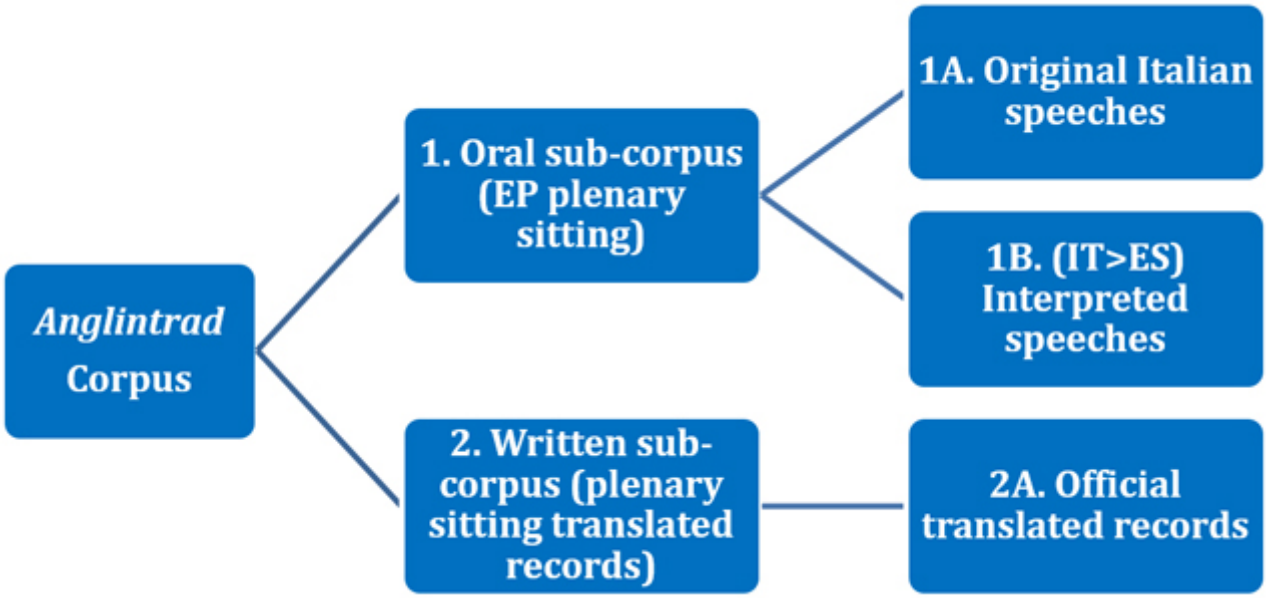


Figure 1. ***Anglintrad*** **structure**

2. The corpus

2.1 Methodology

Following the principles underlying the compilation of EPIC**[3]** and its transcription conventions, the *Anglintrad* corpus was designed to serve a specific purpose: providing a significant amount of data to observe a particular phenomenon, without challenging some basic methodological assumptions which were described by Bendazzoli:

> La creazione di EPIC rappresenta uno dei primi tentativi di superare gran parte degli ostacoli descritti […], in quanto ci si è avvalsi di materiale autentico, omogeneo rispetto a numerose variabili e in quantità sufficienti per essere rappresentativo; lo stesso materiale è stato poi elaborato e reso disponibile in formato elettronico, in modo da poter farne uso a fini di ricerca e didattica attraverso procedure semi-assistite e pertinenti con la linguistica computazionale. […] La scelta del materiale da includere nello studio è stata guidata da molteplici fattori, quali le fonti disponibili, gli strumenti tecnici più idonei alla raccolta, conservazione ed elaborazione del materiale oggetto di studio e le risorse tecnologiche disponibili al momento dell'attivazione del progetto […]. (Bendazzoli 2010: 117)

In the light of the need for data accessibility and above all comparability, the European Parliament plenary sitting was selected as the source of all the materials included in *Anglintrad*. This guarantees not only the authenticity of the original material, one of the main methodological challenges in Corpus-based Interpreting Studies (Shlesinger 1998), but also its homogeneity, since oral data coming from different contexts and settings may compromise the basic principles of the study. The selected texts were all delivered in 2011 in 26 plenary sittings where a total number of 241 items (unmodified English loanwords in the original Italian speeches) were identified.

The unrevised verbatim reports of the original Italian speeches were first scanned as in fast reading in order to detect those containing at least one phenomenon to study; then, the selected texts were analysed and transcribed, following the EPIC transcription conventions**[4]**, and the same procedures were applied to the related Spanish interpreted versions. In the last phase, these text segments were aligned to their official revised translations to allow for an immediate comparison between the three texts (original speech – interpreted version – translated version). A summary of the main methodological steps for the compilation of *Anglintrad* is provided in Fig. 2:

PHASE 1:
Scanning of the unrevised verbatim version of the original Italian speech to detect the presence of unmodified English loanwords (easily identifiable in the Italian text).

PHASE 2:
Listening to the original (Italian) video file and its interpreted (Spanish) version. Transcription following EPIC transcription conventions.

PHASE 3:
Comparison between the interpreted version and the related revised official translated version.

Figure 2. Main methodological steps

2.2 *Anglintrad* characteristics

The comparison between the three main text segments (original Italian speech – interpreted Spanish version – translated Spanish version) was meant to be as immediate and user-friendly as possible, therefore their structure was organized in a spreadsheet.

Every phenomenon identified in the corpus was classified and matched with a set of metadata on the plenary sitting (for example, "*Dibattito 17_01_11*"), a link to the official translated version (for example, "*Resoconto tradotto*"), a link to the related verbatim unrevised version of the original Italian speech (for example, "*Resoconto*"), the specific topic (for example, "*Dichiarazioni del Presidente del Parlamento Europeo sulla situazione in Tunisia*") and the speaker (for example, "*Pier Antonio Panzeri*").

For a quicker comparison between the three versions of the same phenomenon, the structure was divided into three columns: the first one indicates the transcription of the text segment where the phenomenon was identified; the second one includes the transcription of the same text segment in the interpreted version, while the third column includes the official translated version. This layout allows for an immediate comparison between the phenomena, highlighting possible problems and strategies adopted in simultaneous interpreting/translation. This visualization can be easily exploited for didactic purposes in interpreter and translator training.

Another important element to be considered when using these materials for pedagogical objectives is the composition of the corpus itself. As already mentioned, *Anglintrad* includes 241 unmodified English loanwords detected in the speeches delivered by 46 different speakers (32 men and 14 women) during 26 plenary sittings held in 2011; the total number of occurrences delivered by men is 184 and by women is 57. When a loanword was found, a few words before and after the item were transcribed and included in the corpus in order to preserve the meaning of the sentence. Since it is a purpose-specific corpus and given the main research objective, a full transcription of the whole speeches in which a loanword is present is not provided because the research focus is meant to remain within the analysis of this particular phenomenon and the way it is managed by interpreters and translators.

Further information and metadata on the structure of the corpus itself (distribution of phenomena by topic, type of entry, and type of pronunciation in the original Italian speech) is provided in Figures 3-6 below**[5]**:

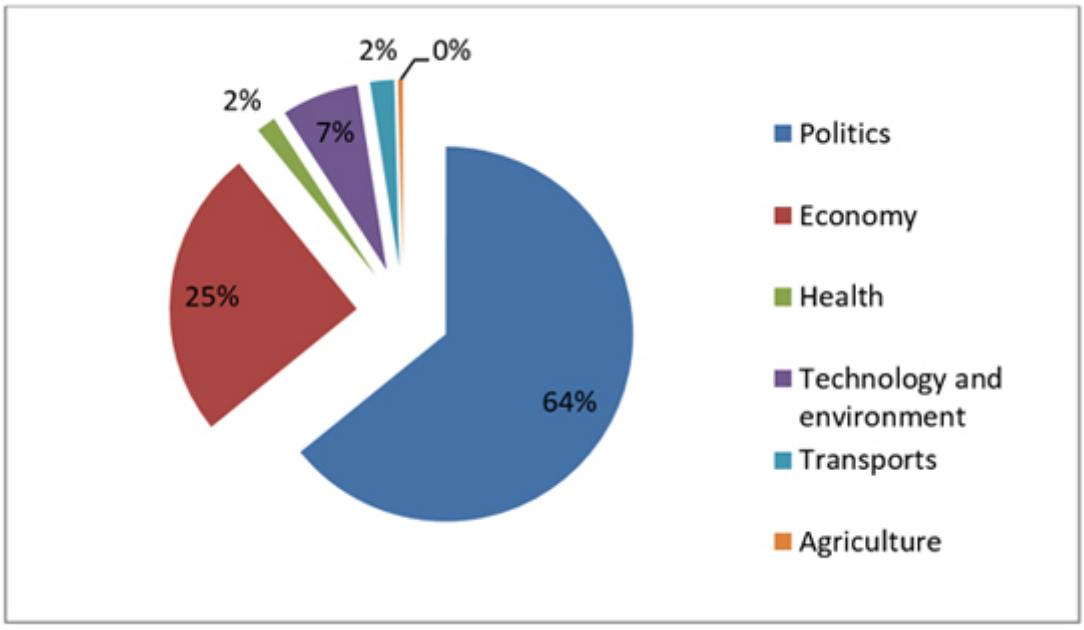


Figure 3. Percentage of loanwords by topic

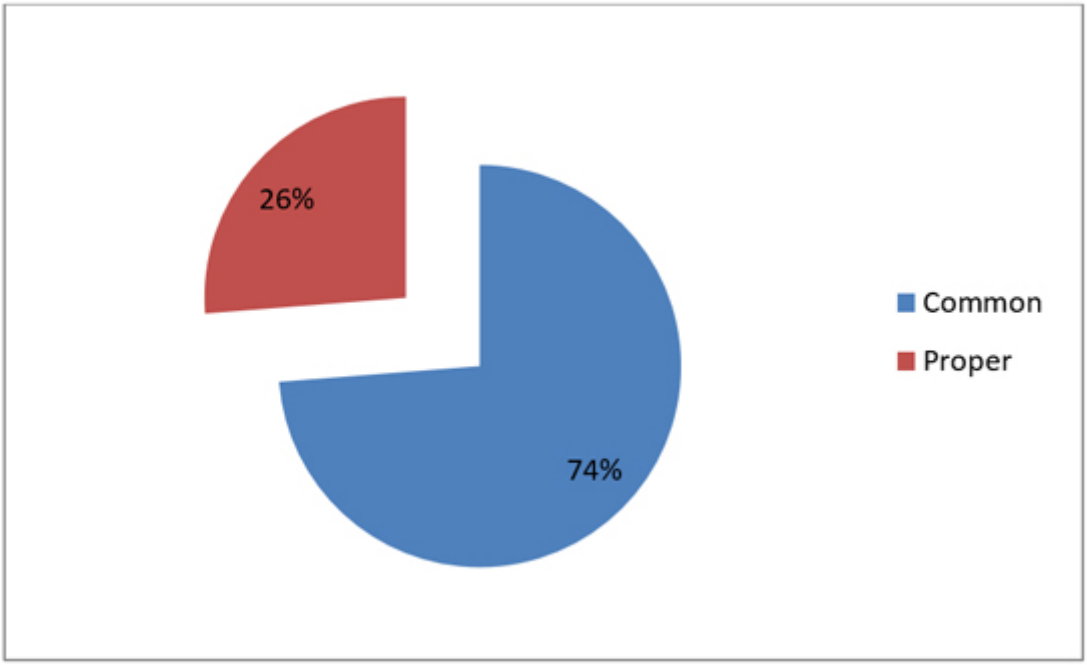


Figure 4. Shares of common and proper items in loanwords

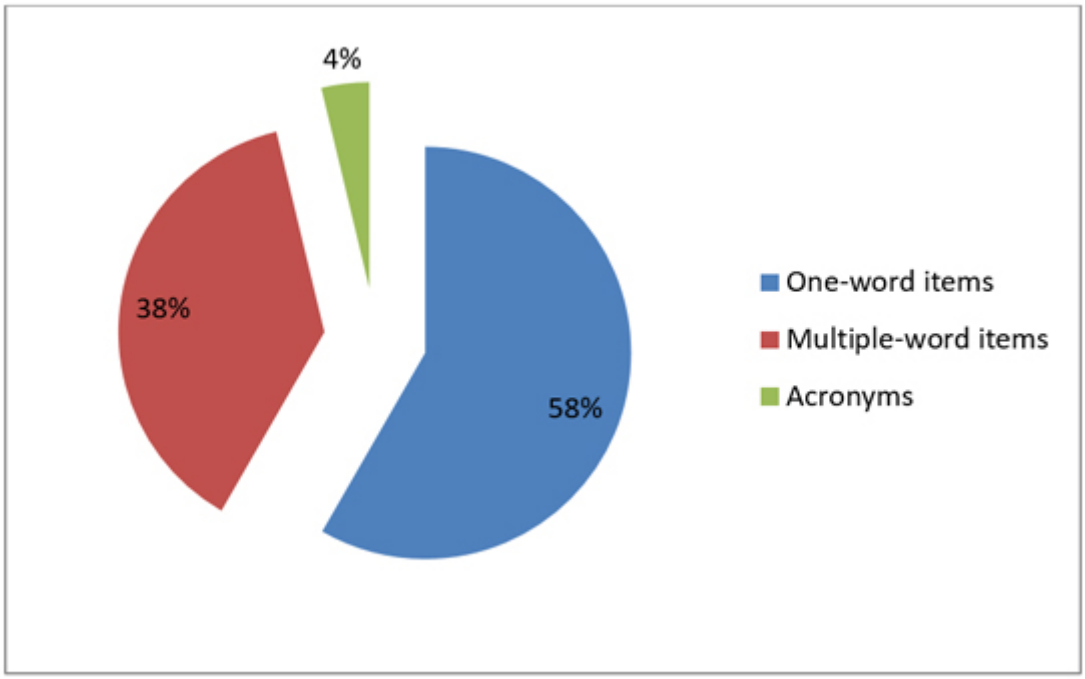


Figure 5. Shares of item types in loanwords

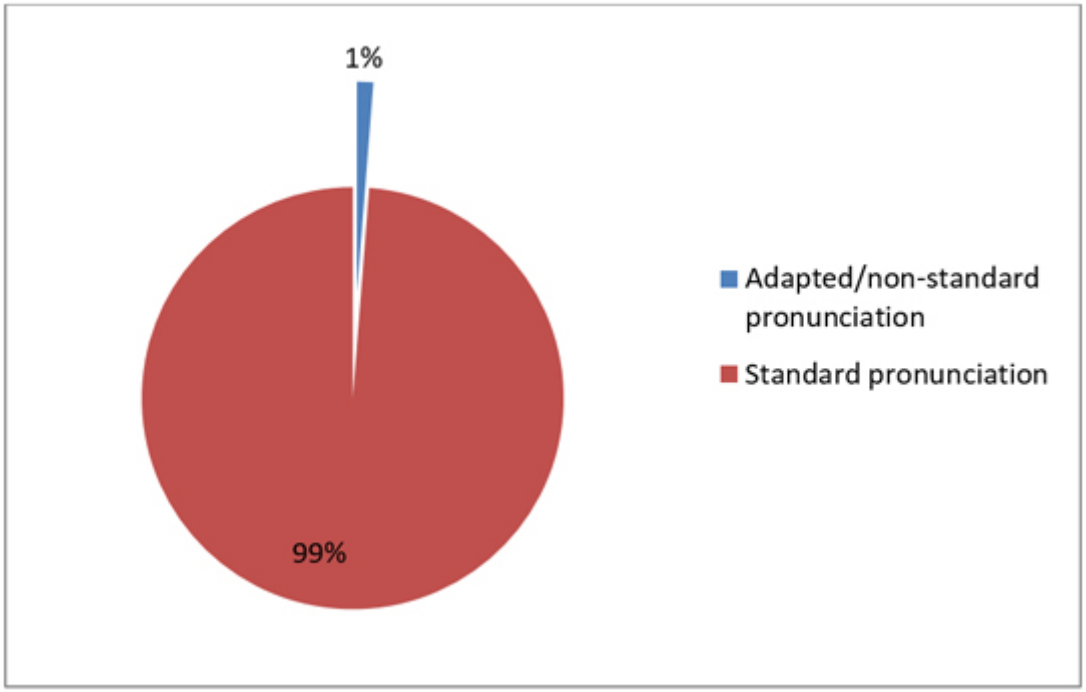


Figure 6. Share of pronunciation type[6] in original Italian speeches

Some weighted percentages**[7]** were also calculated for the number of speakers by gender and the related weighted percentage of phenomena divided by speaker's gender (see Figures 7-8); the number of speakers by political group (S&D – Social and Democrats, EPP – European People's Party, EFD – Europe of Freedom and Democracy, ALDE – Alliance of Liberals and Democrats for Europe) and the related weighted percentage of phenomena by political group (see Figures 9-10):

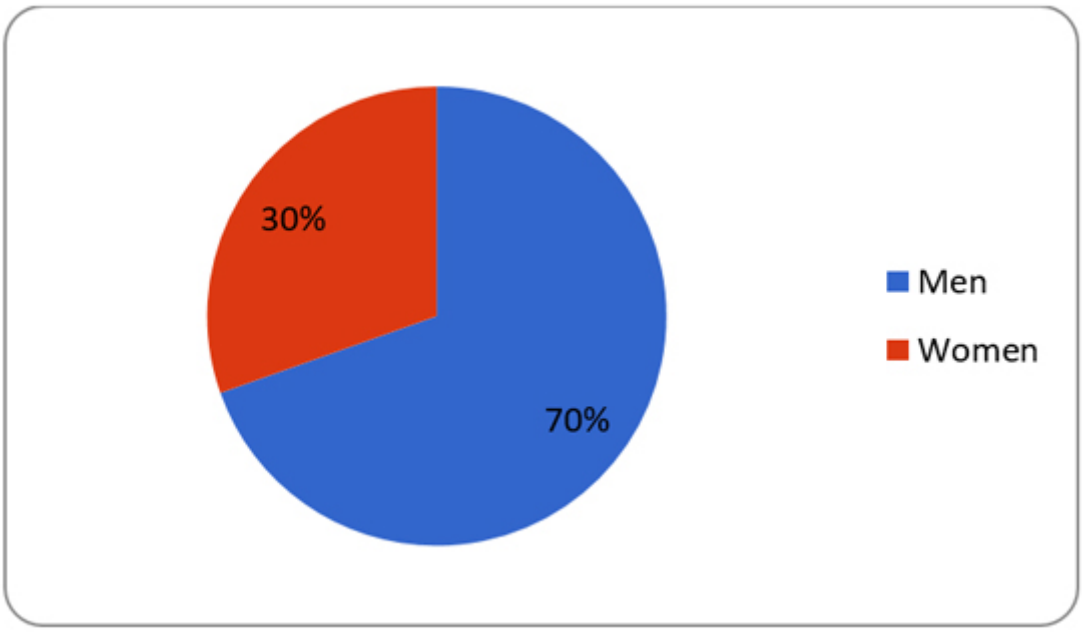


Figure 7. Gender distribution of speakers

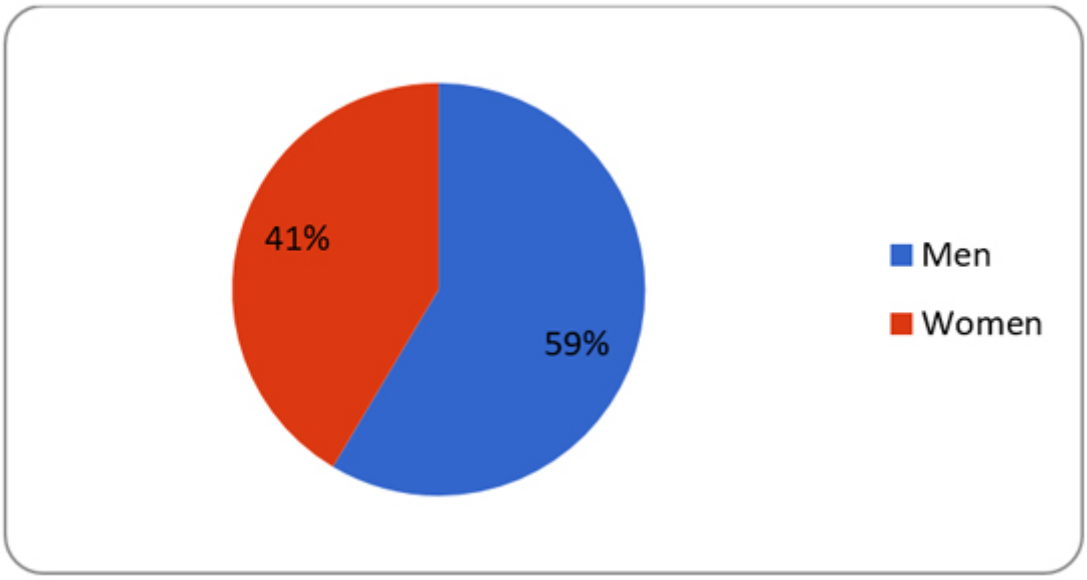


Figure 8. Weighted percentage of loanwords by speaker gender

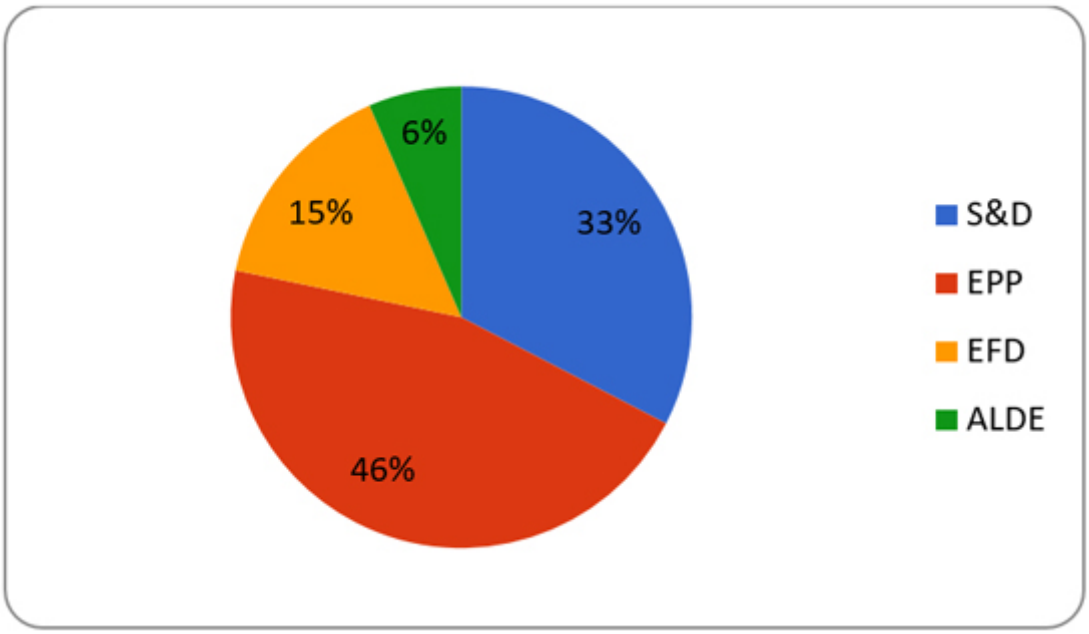


Figure 9. Distribution of speakers over political groups

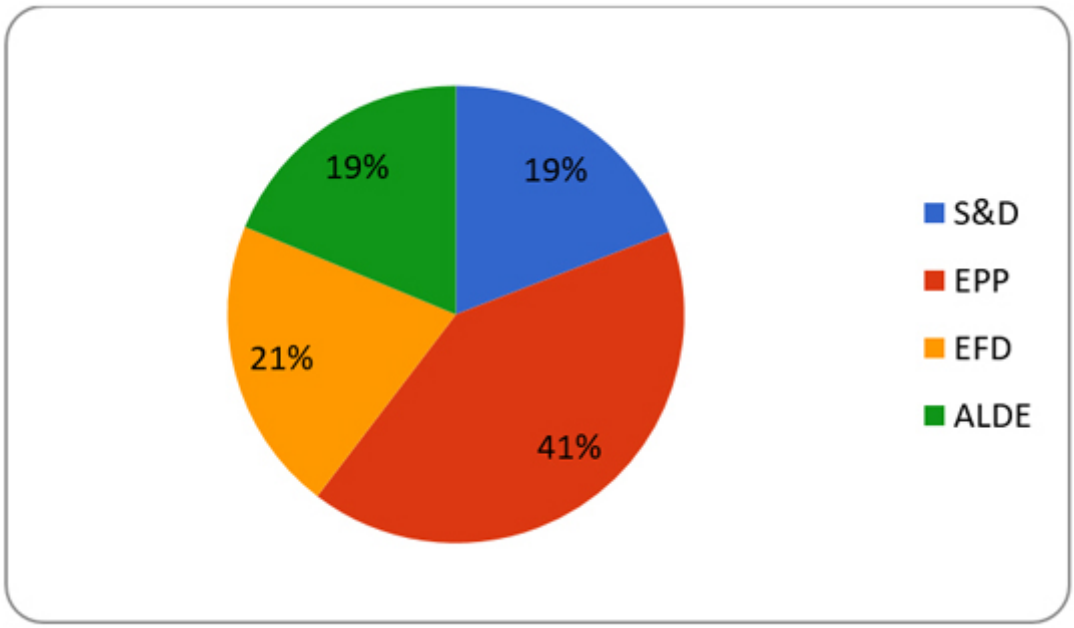


Figure 10. Weighted percentage of loanwords by political group

3. Preliminary results

3.1 Strategies: similarities and differences

The corpus allows for a double-perspective observation of the same phenomenon (unmodified English loanwords in the original Italian speech) because it is intermodal and directly comparable, since the same oral text is aligned to the related interpreted and translated versions. The fact that *Anglintrad* is a purpose-specific corpus (with a particular aim: observing what strategies are applied by simultaneous interpreters and translators to tackle the same potentially challenging lexical item such as unmodified English loanwords in the Italian>Spanish language combination) meets the need to create a user-friendly tool that can be easily exploited for didactic purposes.

A first analysis of the corpus based on the simple observation of unmodified English loanwords highlighted some preliminary results that, despite being far from thorough and complete, can provide an interesting initial overview of the similarities and differences in the strategies used by interpreters and translators dealing with the same lexical item, in the same text and within the same context.

The first step of this empirical observation was an attempt to classify the different strategies detected both in the interpreted and in the translated versions (see Table 1).

List of Strategies	Definition	Italian	Spanish	References
1 Cancellation	The phenomenon is not rendered.	/ […] registro delle **lobby** che dovrebbe essere formato da qui a breve/	XXX	Bakti 2009
2 Exact rendition	The phenomenon is rendered with no modifications.	/[…] la commissione si è sempre schierata a favore del **Made In**/	/[….] a favor del **Made In**/	Wadensjö 2001, Schjoldager 1996
3 Generalization	The communicative intention or the basic concept is rendered in a generic way.	/ […] l'azzeramento dei dazi sui prodotti coreani contro l'innalzamento degli **standard** ambientali e sociali in Corea/	/ […] tendrán demasiadas ventajas, muchas más que los productos europeos/	Al-Khanji *et al.* 2000, Bartlomiejczyk 2006
4 Substitution	The phenomenon is reformulated at a lexical level (use of synonyms) or at a syntactic level.	/[…] i meccanismi di pubblicità il **web** e altre modalità efficaci/	/[…] todos los medios que existen **como...la internet**/	Wadensjö 2001, Straniero Sergio *et al.* 2012
		/l'Europa deve essere in grado di intervenire con misure comuni…ed efficaci per la sicurezza… dell'approvvigionamento alimentare per evitare le forti asimmetrie a-ancora esistenti relative agli **standard** di-di sicurezza tra i prodotti UE ed extra UE grazie/	/Europa debe intervenir con medidas comunes y eficaces…para que haya un surtido alimentario adecuado evitando…las fuertes asimetrías aún existentes ehm... relativas a las **normas** de seguridad...entre productos europeos y no europeos/	
5 Translation	The entry is adapted to the morphological and lexical norms of the target language or the lexicalised equivalent in the target language is used.	/[…] il rimborso del prezzo del biglietto in caso di partenza annullata ritardo superiore alle due ore o **overbooking**/	/[…] que se le reembolse el billete en caso de que se le cancele la salida o un retraso de más de dos horas o cuando haya ehm **sobreventa**/	
6 Expansion	At different levels, the interpreter/translator can make additions to the source text.	/[…] **stress test** […]/	/[…] **pruebas de... aguante / resistencia** […]/	Bartlomiejczyk 2006

Table 1. List of strategies

The first strategy detected in the corpus is **cancellation**, indicating the lack of any type of rendition of the original phenomenon in the target text: at first sight, it may seem that cancellation necessarily entails a partial or complete loss of the original content, but it can also be a strategy activated by interpreters or translators to make the original message clearer, provide better cohesion to the target text or eliminate redundancy if present in the source text (Russo and Rucci 1997). In the case of interpreted texts, this holds even more true since cancellations can:

> […] help guarantee the best possible quality of interpretation under the circumstances. […] In some cases, omissions are deliberate and aimed at economy of expression, ease of listening for the audience and maximum communication between the speaker and audience. (Jones 1998: 139)

The second strategy is **exact rendition**, meaning that the loanword is simply transposed into the target language without any type of modification. In the light of the different mechanisms that these cognate languages (Italian and Spanish) have to integrate new unmodified loanwords (Tonin 2010), this strategy may be regarded as potentially complex to be applied correctly. However, the typical features of this setting (the European Parliament plenary sitting) and its microlanguage (Bertozzi 2016) allow for a purpose-specific use of both oral and written language, since all participants share the same knowledge and background: that is why a "non-domestication" strategy may be perfectly acceptable in this setting:

> L'unico soggetto che potrebbe discostarsi dal gruppo è lo stesso interprete, in quanto difficilmente avrebbe la possibilità di condividere lo stesso livello di esperienza e preparazione degli altri partecipanti, pur preparandosi adeguatamente all'incarico assegnato. In questo caso l'interprete prediligerebbe il più possibile un uso tecnico e specifico della lingua; eventuali lacune sarebbero generalmente compensate dalla conoscenza degli ascoltatori. (Bendazzoli 2010: 151)

The third strategy detected in the corpus is **generalization**, where the communicative intention or the basic concept of the source text is rendered in a generic way in the target text. This category also includes:

> […] l'utilizzo di acronimi e il ricorso alla deissi, utilizzata in sostituzione di porzioni di testo più lunghe, grazie alle conoscenze che l'interprete condivide con oratore e pubblico. (Voncina 2009: 28)

This technique has always been widely used and studied both in Interpreting and Translation Studies; more specifically, in the case of interpreting, Gile (1995) included generalization among the so-called "preventive and reformulation tactics" consisting of 'replacing a segment with a superordinate term or a more general speech segment' (Bartlomiejczyk 2006: 152).

Substitution is the fourth type of strategy identified in the corpus, meaning that the phenomenon is reformulated at a lexical level (in other words, with the use of synonyms) or at a syntactic level. This macro-category includes a set of sub-strategies such as morpho-syntactic transformation, chunking (Seleskovitch and Lederer 1989), permutation or the re-arrangement of elements within the same sentence (Pippa and Russo 2002) and paraphrasing. Restricting the scope to interpreting, this strategy can be particularly demanding in terms of cognitive load and lexical retrieval capacity since, in some cases, this type of rendition is far longer and more complex than the original message, with a subsequent lengthening in the interpreter's *décalage* and possible carry-over effects in the following segments. That is why it must not come as a surprise that 'experts did more than twice as much lexical elaboration than novices' (Setton and Motta 2008: 217).

The fifth strategy is **translation**, where the phenomenon is adapted to the morphological and lexical norms of the target language or where the lexicalised equivalent in the target language is used. If, on the one hand, the Italian language has always tended to integrate unmodified loanwords (possibly modifying only their phonetic level), on the other hand the Spanish language has a more restrictive approach and tends to use target terms more frequently (Tonin 2010). Many examples such as "*budget - presupuesto*", "*road map – hoja de ruta*" or "*bond – bono*" can be found in the corpus.

The sixth and last strategy is **expansion**, where the interpreter/translator makes additions to the source text at different levels. In interpreting, this phenomenon has also been called "addition":

> Addition is treated as a strategy when the interpreter decides to add, by way of explanation, something the original speaker did not say because the interpreter thinks the interpretation may otherwise not be clear for the audience (e.g. due to discrepancies between the source- and target-language cultures).(Bartlomiejczyk 2006: 160)

This holds true also for translation, where expansion can be used for discourse-planning purposes, or to provide better cohesion to the target text.

After direct comparison between the three versions of the same phenomenon and the identification of a set of strategies (which, far from being exhaustive, can however provide a necessary attempt to classify the different strategies adopted for didactic purposes), the next step entailed the subdivision of the strategies adopted by interpreters and translators into two main macro-categories: same and different strategies, where the first ones include marked similarities (at a lexical, pragmatic level, and so on) between the interpreted and the translated renditions, while the second ones indicate how the interpreter and the translator facing the same linguistic phenomenon can adopt different strategies (cancellation, exact rendition, generalization, substitution, translation, expansion) (see Figures 11-12). This type of classification proved to be the most suitable for didactic purposes, where the need to simplify this structure as much as possible and therefore the need to provide a user-friendly tool for interpreting and translation trainees is crucial:

Figure 11. Example of identical strategies

The example in Figure 11 shows that the interpreter reformulated the segment containing the loanword (*standard di sicurezza*) at a lexical level (*normas de seguridad*), which can be classified as a substitution (see fig. 12); the use of this strategy in simultaneous interpreting is particularly frequent and reformulation is often associated with the activity of interpreting itself:

> L'abitudine alla riformulazione, a una maggiore flessibilità lessicale può trasformarsi in una strategia automatizzata che consente di distribuire al meglio le proprie risorse per prevenire una resa insoddisfacente imputabile a una cattiva suddivisione delle stesse. (Riccardi 1999: 172)

More specifically, with regard to the example above (Figure 11), one can hypothesize that the interpreter tried to retrieve the same word in Spanish and the latter may have not been immediately available in his/her memory (Gran 1992), as one could assume given the presence of a filled pause (*ehm…*) just before this segment (Ahrens 2002); therefore, due to time constraints, the interpreter may have tried to find a possible strategy to render this potentially challenging phenomenon (an unmodified loanword from a third language that is not included in the pair being activated in simultaneous mode) by reformulating the source segment. Interestingly, despite the many obvious differences characterizing translation and interpreting activities, the same strategy (substitution) was activated by the translator as well (*normas de seguridad*). This may suggest that what may seem to be an "emergency strategy" in interpreting (reformulation as a consequence of difficulties in retrieving the right word/segment) can actually be a specifically-targeted strategy as such in translation: as a matter of fact, the segment "*normas de seguridad*" is particularly frequent in Eurlex**[8]**, so this may prove that the translator was provided with specific terminological guidelines in advance (which may also apply to interpreters, but the simultaneous mode does not always allow for an immediate retrieval of single specific terms, even if provided in advance).

An example of different strategies activated by interpreters and translators in the corpus is provided in figure 12:

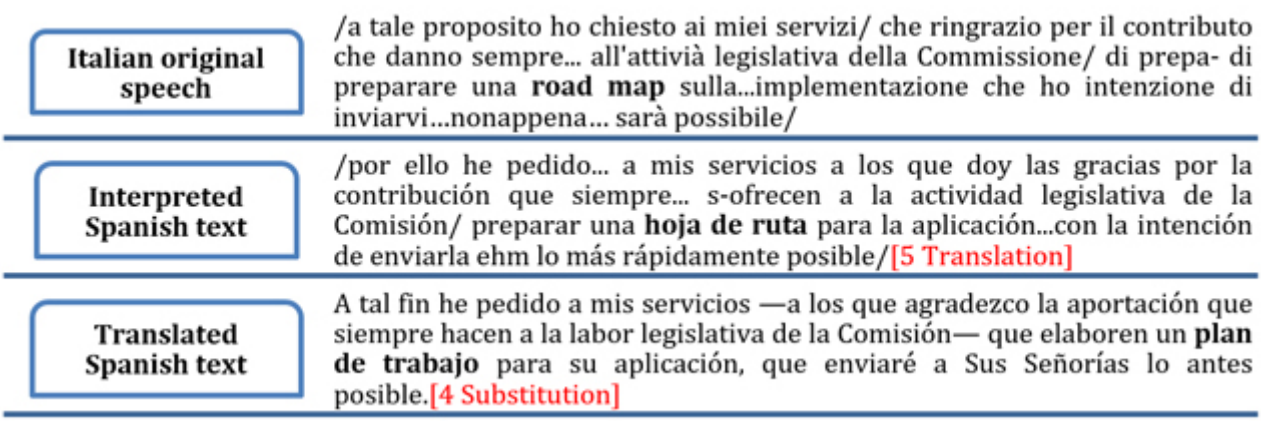


Figure 12. Example of different strategies

In this case, the original speaker is making use of an unmodified loanword (*road map*) that is becoming more and more common in the Italian language, especially in the press and in the political domain. Given the importance of an in-depth analysis for each type of loanword, its main characteristics and use in modern Italian, every phenomenon identified in the Italian sub-corpus was provided with a specific terminological sheet (an example is provided in Table 2) indicating its grammatical category, gender and number, the related original word in English, its definition taken from main modern Italian dictionaries, the use of the linguistic phenomenon in context (from the Lexis Nexis Database**[9]**), the year of first appearance in dictionaries (where reported), its further productivity in Italian (if any), any indications on pronunciation and some information on the history of the loanword in Italian (whether it is a neologism, it is reported as "anglicism" in the dictionaries, it is present in previous editions of the same dictionary or it is part of a sectoral language):

Lessema	ROAD MAP
Categoria grammaticale	lessema ingl. (propr. «carta stradale»), usato in ital. come sost. femm.
Genere	femm.
Numero	invar. (Gabrielli); Treccani ammette il plur. road maps ‹... mäps›.
Derivazione inglese (Oed)	noun; 1A map, especially one designed for motorists, showing the roads of a country or area. 2A plan or strategy intended to achieve a particular goal: "a road map for peace in the region".
Fonti lessicografiche /terminologiche italiane	VOCABOLARIO TRECCANI: 1. Spec. nel linguaggio giornalistico, piano diplomatico e strategico accuratamente programmato, e da realizzarsi in diverse tappe, in vista del raggiungimento di uno specifico obiettivo, spec. con riferimento al conflitto tra israeliani e palestinesi. 2. estens. Tabella di marcia, programma di lavoro e sim.: attenersi scrupolosamente alla road map fissata. DIZIONARIO GABRIELLI: Piano, progetto dettagliato, scandito a tappe come una tabella di marcia, in vista di un obiettivo da perseguire.
Contesti	Il piano di pace del Quartetto Usa-Ue-Onu-Russia, la cosiddetta *'road map'*, e' stato ufficialmente presentato questo pomeriggio al nuovo premier palestinese Mahmud Abbas a Ramallah (Ansa 2003 - Database Lexis Nexis). Toccherà al Quartetto, cioè ai quattro mediatori internazionali (Stati Uniti, Russia, Unione europea e Nazioni Unite) che hanno redatto la *road map*, valutare i progressi nell'attuazione del piano (La Stampa 2003 - Database Lexis Nexis). Il nuovo capo dell'Anp Abu Mazen ha detto oggi che i palestinesi sono pronti a attuare gli impegni assunti nella *Road Map*, il percorso di pace delineato due anni fa dal Quartetto Usa-Onu-Ue-Russia (Ansa 2005 - Database Lexis Nexis). Se la divisione destra/sinistra ha ancora un senso, e si rimprovera alla *road map* di Monti di aver seguito politiche sbilanciate nell'una o nell'altra direzione, in una coalizione destra-sinistra rimproveri del genere non sono evitabili e segnalano che la *road map* funzionerebbe meglio se avesse alle sue spalle una maggioranza politicamente coerente (Corriere della Sera 2012 - Database Lexis Nexis). In quella sede, sono emerse, nei tavoli di lavoro, le varie proposte della *Road Map* in ausilio ed in funzione della legge di stabilità 2016 (Italia Oggi 2015 - Database Lexis Nexis).
Anno	2003 (Treccani).
Produttivita' del lessema/ulteriori apporti dall'inglese	La locuzione nasce in un contesto ben specifico, quello del conflitto israelo-palestinese (piano diplomatico e strategico accuratamente programmato, e da realizzarsi in diverse tappe, in vista del raggiungimento di uno specifico obiettivo, spec. con riferimento al conflitto tra israeliani e palestinesi) e si estende in seguito alla seconda accezione (Treccani), ad oggi molto frequente: tabella di marcia, programma di lavoro.
Indicazione di pronuncia	‹róud mäp›, road maps ‹... mäps› (Treccani). Non indicata in Gabrielli.
Riferimenti	**http://www.treccani.it/vocabolario/road-map/** (19/02/16) **http://dizionari.repubblica.it/Italiano/R/roadmap.php** (19/02/16) **http://www.oxforddictionaries.com/definition/english/road-map?q=road+map** (19/02/16) **https://it.wiktionary.org/wiki/road_map** (19/02/16).
Note	prestito linguistico dall'inglese *road map*, entrato nel linguaggio italiano inizialmente tramite il gergo giornalistico per riferirsi al processo di pace israelo-palestinese (Treccani, Wikizionario).
Carattere neologico	1) PRESENZA NEI DIZIONARI DI LINGUA GENERALE: sì (Treccani 2003, Gabrielli). Non indicato da De Mauro né Sabatini Coletti. Il Dizionario De Agostini 1995 e lo Zingarelli 1970 non lo riportano. 2) SEGNALATO COME ANGLICISMO: sì, da Treccani. Non segnalato da Gabrielli. 3) PRESENZA INDICAZIONE DI PRONUNCIA: solo Treccani la

	riporta. 4) LINGUAGGIO SETTORIALE/LINGUA GENERALE: il lessema scaturisce dal linguaggio giornalistico, con particolare riferimento al conflitto israelo-palestinese e solo successivamente si estende al linguaggio generale nella sua accezione più ampia di tabella di marcia, programma di lavoro (Treccani).

Table 2. Example of terminological sheet

In the specific case illustrated in Figure 12, the use of the loanword "*road map*", which entered the Italian vocabulary through the journalistic language with reference to the Israeli-Palestinian conflict, is becoming more and more frequent, thus potentially affecting the way interpreters may deal with this phenomenon: as a matter of fact, interpreters tend to rely on automatic mechanisms to render the most frequent linguistic features (such as loanwords). This may be one of the reasons why the interpreter did not hesitate in using a translation strategy in this case ("*hoja de ruta*" is the semantic equivalent of "road map"); the translator did not opt for the same strategy, relying on a substitution ("*plan de trabajo*"), which may seem to be more frequent in interpreting (due to time constraints and the difficulties in retrieving the exact word, thus potentially leading to a reformulation). In this case, the translator's aim was making the target text more recipient-oriented and clearer from a linguistic point of view.

Another example of same strategies used by interpreters and translators in the corpus is provided in Figure 13:

Italian original speech	/alcuni diritti positivi vengono tutelati come quelli dei portatori di **handicap**/
Interpreted Spanish text	/hay derechos positivos bien plasmados y tutelados y amparados sobre todo entonces los **discapacitados o personas con movilidad reducida**/[6 Expansion]
Translated Spanish text	Algunos derechos están protegidos, como los de **personas con discapacidad o con movilidad reducida**.[6 Expansion]

Figure 13. Example of same strategies

The loanword *handicap* (and the related expression *portatori di handicap*) has a long tradition in the Italian vocabulary (the first occurrences in the main Italian dictionaries date back to the late 19th century); the same applies for the Spanish language, but with a difference: the entry in the *Diccionario de la Real Academia Española***[10]** is *hándicap* (with acute accent) and, in the *Diccionario Clave***[11]**, *handicap* is in italics, since it is classified as a foreign word. The *Diccionario Panhispánico de Dudas***[12]** suggests the use of *discapacitado* or *minusválido* instead of the unnecessary anglicism *handicapado*. The choice made by the interpreter and the translator in this case (Figure 13) is particularly interesting because they both rely on an expansion, a strategy requiring additional efforts, especially in the simultaneous mode (Bartlomiejczyk 2006). It appears quite clearly that both interpreters and translators are particularly sensitive to the "politically correctness" issues inherent in language and one could assume that, within the European institutions, they are provided with guidelines on how to render/translate these potentially challenging phenomena: this could be one of the reasons why they both opted for an expansion of the original text, even if there was no need to further clarify the source message and despite the fact that the previous segment might have been particularly difficult to render in simultaneous mode.

4. Conclusions

In this paper the methodological steps undertaken to create a bilingual (Italian > Spanish) intermodal (simultaneous interpreting and written translation) corpus for pedagogical purposes have been presented. The *Anglintrad* corpus is being built to explore the strategies used by interpreters and translators when dealing with unmodified English loanwords in the Italian source text. An easy-to-use classification of interpreting/translation strategies along with convenient parallel display of both source and target texts have been designed and can be exploited in interpreter and translation training.

A thorough analysis of the whole corpus was still beyond the scope of the present work. However, a preliminary data observation highlighted that in some cases translation and simultaneous interpreting are much closer than could be expected in terms of strategies applied to face the same problem within the same context and setting (see Figures 11 and 13). This may be due to the fact that both interpreters and translators within the European Parliament share a similar background, a demanding specialized training and the use of standardized terminology for certain terms is highly recommended by the DG Translation and Interpreting.

In other cases observed so far (see Figure 12), the strategies adopted by interpreters and translators can vary considerably due to a number of factors that are not only linked to the different time constraints but also to the different purposes and recipients of the interpreted and the translated renditions.

This dual approach in the observation of the same linguistic phenomenon provides a valuable opportunity entailing resourceful teaching applications for interpreter and translator training and practice. More specifically, in addition to the corpus under construction, the *Anglintrad* project includes the creation of a platform containing useful material for didactic purposes; this platform is currently being developed and will soon make the following material available to interpreting and translation trainees and teachers: first, the bilingual intermodal corpus (Italian original text – Spanish interpreted rendition – Spanish translated version) as described in Section 2, containing additional information on the speaker (name and surname, political group, sex), the type of source text (topic, delivery speed, type of delivery – impromptu or read) and the type of phenomenon detected in the source text (one-word or

multiple-word anglicism, proper or common item); second, a terminological sheet for each phenomenon in the Italian sub-corpus (see Table 2), containing an in-depth analysis of the loanword and its history/use in the Italian language (frequency of use, definitions, contexts, specific domains, and so on); finally, a user-friendly classification of the strategies adopted by interpreters and translators (see Table 1), allowing for direct comparison between the two (same/different strategies).

This twofold (interpreting vs translation) perspective has already been hypothesized by some scholars. It is in particular worth mentioning the article by Viezzi (1993) in this context, in which written translation and simultaneous interpreting are contrasted in a case study and the work by Padilla Benítez et al. (1999), who apply the principles of cognitive theory to the two disciplines. However, the same approach has never been used to study a specific linguistic feature: that is the reason why the creation of an open-access platform for the analysis and comparison of the strategies adopted by interpreters and translators, as well as the main challenges involved, can be particularly beneficial for didactic purposes and can provide new insights based on different perspectives and strategies, bearing in mind that each discipline can learn something from the other.

A preliminary analysis of the corpus suggests that the same strategies are used more often than one might expect: this can serve as a starting point for a new approach in translation and interpreting training, providing a platform that collects a number of genuine examples from a real setting and some useful tools (such as the terminological sheets) to raise awareness among trainees on the issue of unmodified English loanwords in Italian, bearing in mind that a good target text (regardless of the translation or interpreting mode) is intrinsically linked to a deep knowledge of the most important emerging trends and features of the source language.

Finally, the same purpose-specific approach can also be applied to the study of other particularly challenging linguistic items and other language combinations as well, paving the way for future research projects and applications.

Appendix

For each chart included in the paper, the related raw frequencies are reported in the frequency bar charts below:

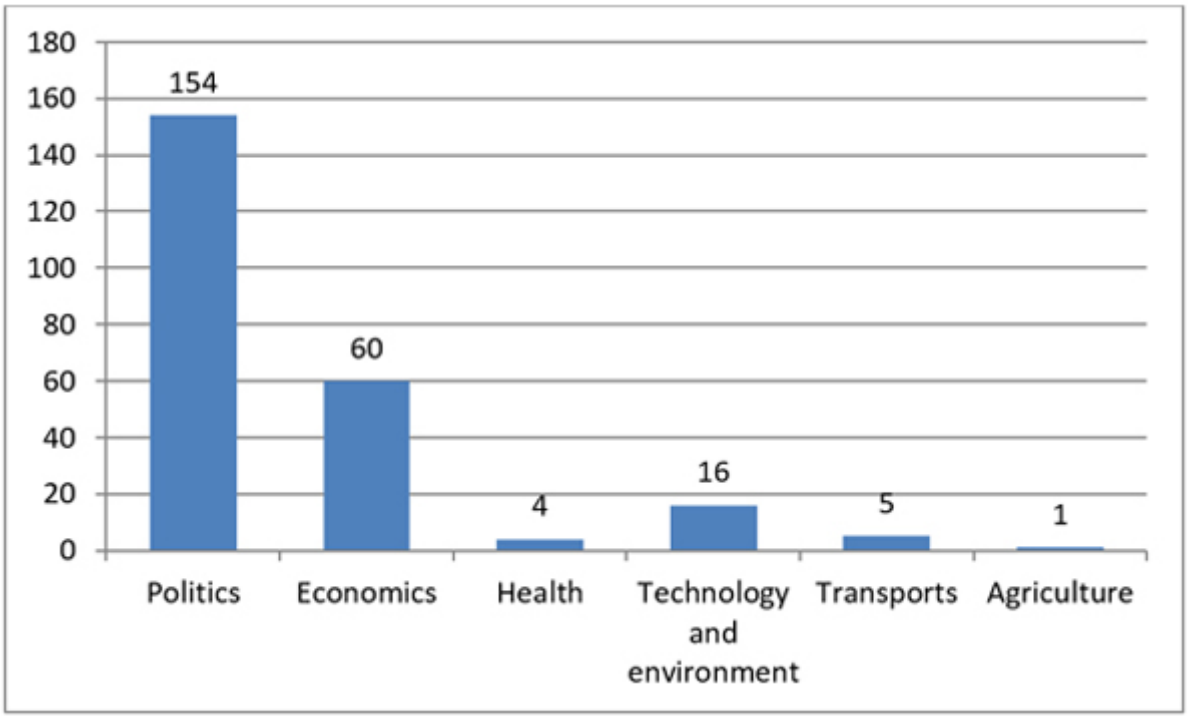


Figure A1. Frequencies of loanwords per topic

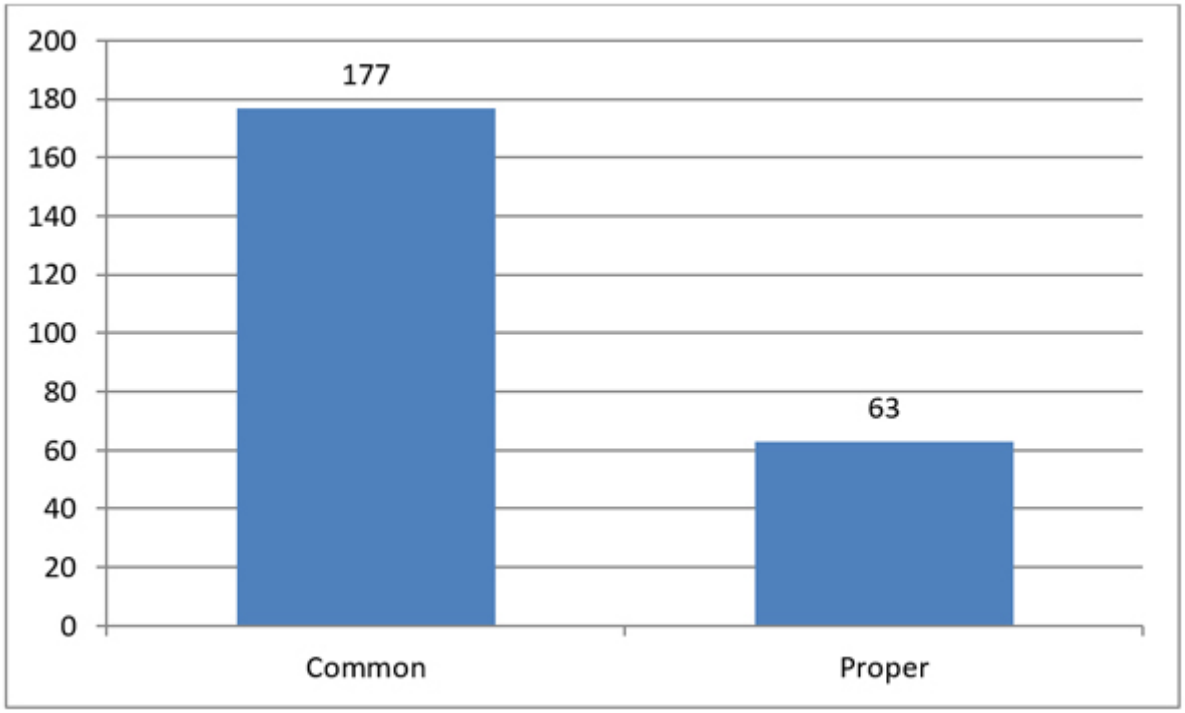


Figure A2. Frequencies of loanwords: common and proper items

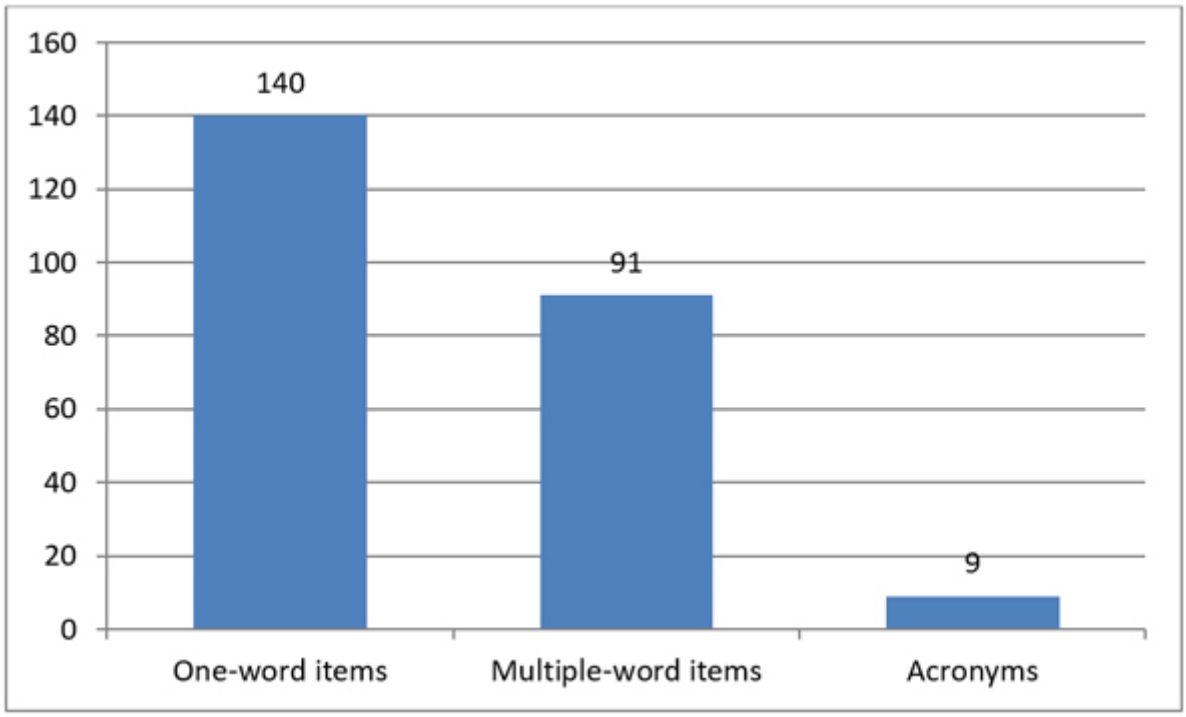


Figure A3. Frequencies of loanwords per type of item

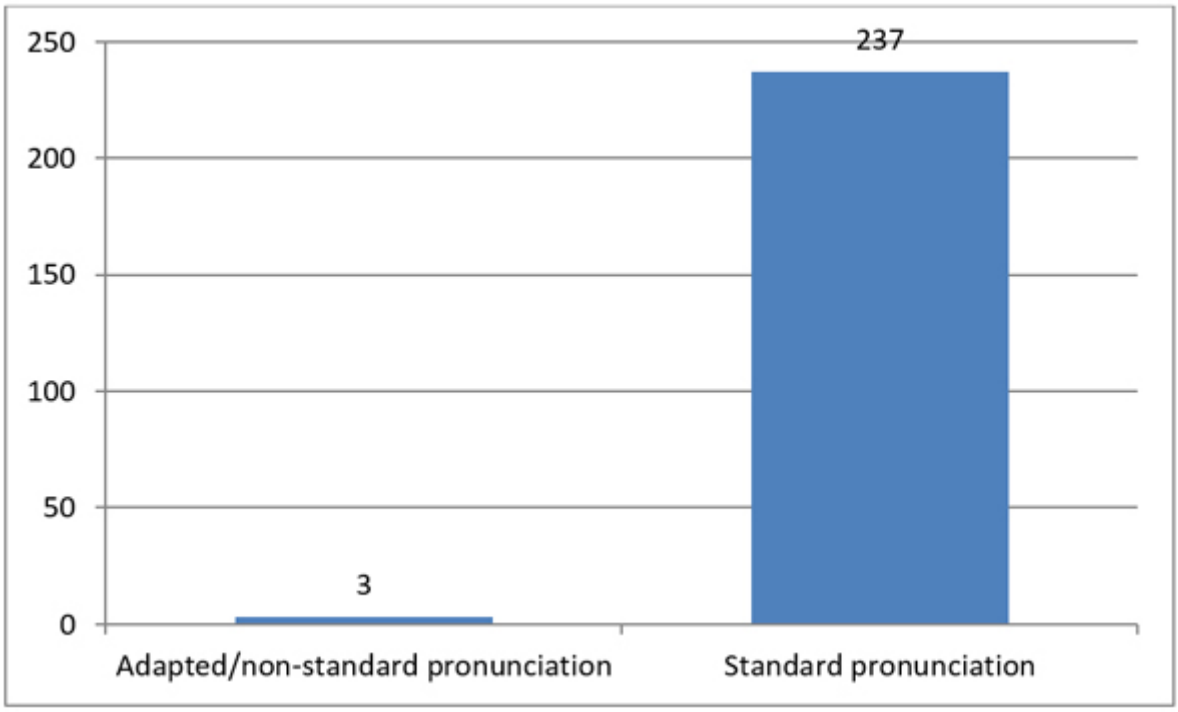


Figure A4. Type of loanword pronunciation in the original Italian speeches

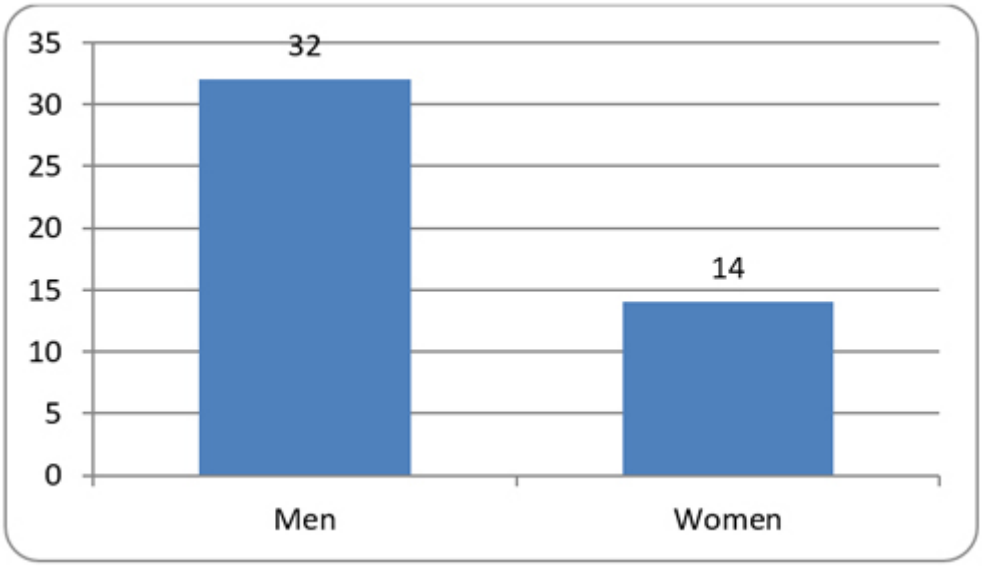


Figure A5. Number of speakers per gender

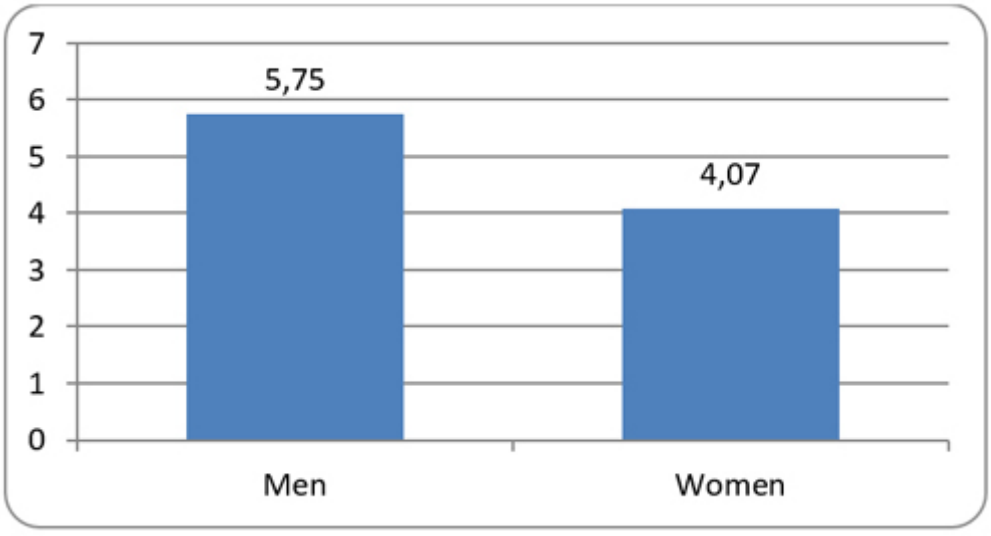


Figure A6. Weighted percentage of loanwords per speaker gender

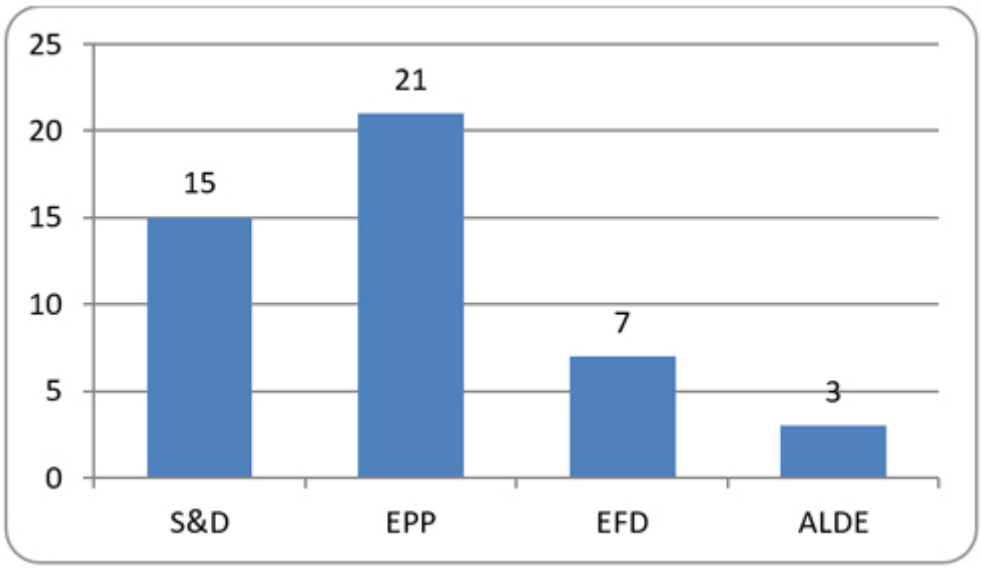


Figure A7. Number of speakers per political group

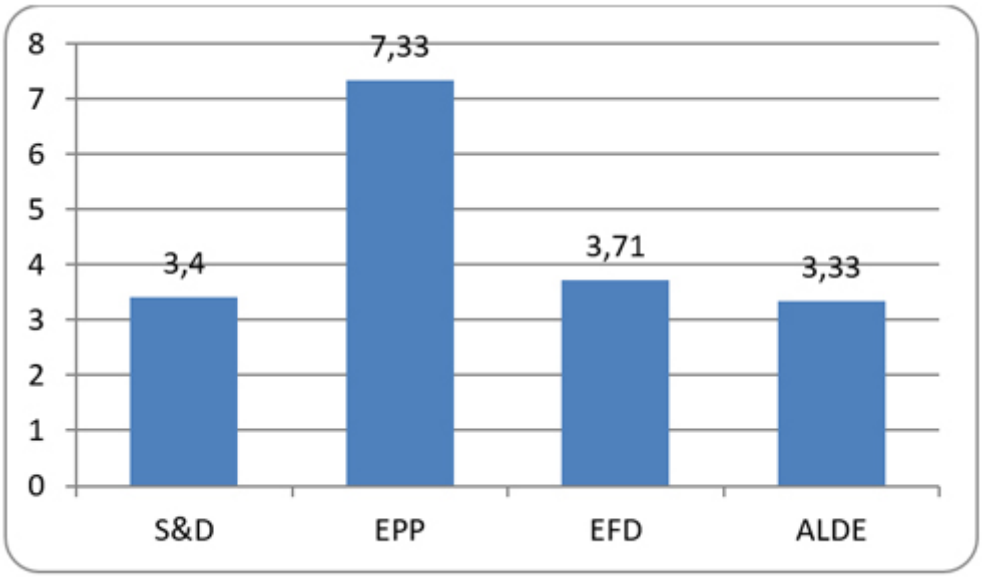


Figure A8. Weighted percentage of loanwords per political group

References

Ahrens, Barbara (2002) "The Correlation between Verbal and Nonverbal Elements in SI" in *Perspectives on Interpreting*, Garzone, Giuliana, Peter Mead and Maurizio Viezzi (eds), Bologna, CLUEB: 37–46.

Al-Khanji, Raja, Said El-Shiyab and Riyadh Hussein (2000) "On the Use of Compensatory Strategies in Simultaneous Interpretation", *Meta* 45, no. 3: 548–57.

Baker, Mona (1993) "Corpus Linguistics and Translation Studies: Implications and Applications" in *Text and Technology: In honour of John Sinclair*, Mona Baker, Gill Francis and Elena Tognini Bonelli (eds), Amsterdam, John Benjamins: 233–50.

Bakti, Maria (2009) "Speech Disfluencies in Simultaneous Interpretation" in *Selected Papers of the CETRA Research Seminar in Translation Studies 2008*, Dries de Crom (ed.), **http://www.arts.kuleuven.be/cetra/papers** (accessed 20 September 2016).

Bartlomiejczyk, Magdalena (2006) "Strategies of Simultaneous Interpreting and Directionality", *Interpreting* 8, no. 2: 149–74.

Bendazzoli, Claudio (2010a) *Il corpus DIRSI: creazione e sviluppo di un corpus elettronico per lo studio della direzionalità in interpretazione simultanea*. PhD diss., University of Bologna.

---- (2010b) *Corpora e interpretazione simultanea*, Bologna, Asterisco. **http://amsacta.unibo.it/view/series/Alma-DL=2E_Saggi.html** (accessed 20 September 2016).

Bendazzoli, Claudio, and Annalisa Sandrelli (2005/2007) "An Approach to Corpus-based Interpreting Studies: Developing EPIC (European Parliament Intepreting Corpus)" in *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation - Saarbrücken 2-6 May 2005*, Heidrun Gerzymisch-Arbogast and Sandra Nauert (eds), **http://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html** (accessed 20 September 2016).

Bendazzoli, Claudio, Cristina Monti, Annalisa Sandrelli, Mariachiara Russo, Marco Baroni, Silvia Bernardini, Gabriele Mack, Elio Ballardini, and Peter Mead (2004) “Towards the Creation of an Electronic Corpus to Study Directionality in Simultaneous Interpreting” in *Compiling and Processing Spoken Language Corpora. LREC 2004 Satellite Workshop IV International Conference on Language Resources and Evaluation*, Nelleke Oostdijk, Gjiert Kristoffersen, and Geoffrey Sampson (eds), Paris, ELRA: 33–9.

Bertozzi, Michela (2016) “Distinctive Features of Orality in a Microlanguage: The Italian Language in the Plenary Sessions of the European Parliament. Some Preliminary Observations”, *MonTI Special Issue*, no. 3: 339–66.

Bombi, Raffaella (2005) *La linguistica del contatto. Tipologie di anglicismi nell’italiano contemporaneo e riflessi metalinguistici* [Contact linguistics. Typology of anglicisms in current Italian and some metalinguistic considerations], Rome, Il Calamo.

Cencini, Marco, and Guy Aston (2002) “Resurrecting the Corp(us/se): Towards an Encoding Standard for Interpreting Data” in *Interpreting in the 21st century: Challenges and Opportunities*, Giuliana Garzone and Maurizio Viezzi (eds), Amsterdam, John Benjamins: 47–62.

Furiassi, Cristiano (2010) *False Anglicisms in Italian*, Monza, Polimetrica.

Gile, Daniel (1995) *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam, John Benjamins.

Gran, Laura (1992) *Aspetti dell’organizzazione cerebrale del linguaggio: dal monolinguismo all’interpretazione simultanea*, Udine, Campanotto.

Gusmani, Roberto (1981) *Saggi sull’interferenza linguistica*, Firenze, Le Lettere.

House, Juliane, Bernd Meyer, and Thomas Schmidt (2012) “CoSi – A Corpus of Consecutive and Simultaneous Interpreting” in *Multilingual Corpora and Multilingual Corpus Analysis*, Thomas Schmidt, and Kai Worner (eds), Amsterdam, John Benjamins: 295–304.

Iglesias Fernández, Emilia (2007) *La didáctica de la interpretación de conferencias. Teoría y práctica,* Granada, Comares.

Jones, Roderik (1998) *Conference Interpreting Explained*, Manchester, St. Jerome.

Kalina, Sylvia (1998) *Strategische Prozesse beim Dolmetschen: Theoretische Grundlagen, empirische Fallstudien, didaktische Konsequenzen* [Strategic processes in interpreting: Theoretical principles, empirical field studies and their implications for teaching], Tübingen, Gunther Narr.

Marzocchi, Carlo (2007) “Translation — Transcript — Interpretation. Notes on the European Parliament Verbatim Report of Proceedings”, *Across Languages and Cultures* 8, no. 2: 249–54.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli, and Mariachiara Russo (2005) “Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus)”, *Meta* 50, no. 4. **http://id.erudit.org/iderudit/019850ar** (accessed 20 September 2016).

Padilla Benítez, Presentación, Maria Teresa Bajo and Francisca Padilla Adamuz (1999) “Proposal for a Cognitive Theory of Translation and Interpreting. A Methodology for Future Empirical Research”, *The Interpreters’ Newsletter* 9: 61–78.

Petite, Christelle (2005) “Evidence of Repair Mechanisms in Simultaneous Interpreting: A Corpus-based Analysis”, *Interpreting* 7, no. 1: 27–49.

Pippa, Salvador and Russo Maria Chiara (2002) "Aptitude for Conference Interpreting: A Proposal for a Testing Methodology Based on Paraphrase", in *Interpreting in the 21st Century. Challenges and Opportunities*, Garzone Giuliana and Maurizio Viezzi (eds), Amsterdam, John Benjamins: 245–56.

Pöchhacker, Franz (1994) “Quality Assurance in Simultaneous Interpreting”, in *Teaching Translation and Interpreting 2: Insights, Aims and Visions. Papers from the Second Language International Conference, Elsinore, Denmark 4-6 June 1993*, Cay Dollerup and Annette Lindegaard (eds), Amsterdam, John Benjamins: 232–42.

Riccardi, Alessandra (1999) “Interpretazione simultanea: strategie generali e specifiche”, in *Interpretazione simultanea e consecutiva: problemi teorici e metodologie didattiche*, Falbo, Caterina, Maria Chiara Russo and Francesco Straniero Sergio (eds), Milan: Hoepli: 161–74.

Russo, Maria Chiara, and Rucci Marco (1997) “Verso una classificazione degli errori nella simultanea dallo spagnolo in italiano" [Towards a classification of errors in Spanish-Italian simultaneous interpreting], in *Nuovi orientamenti negli studi sull’interpretazione*, Gran Laura and Alessandra Riccardi (eds), Padova, Università degli Studi di Trieste: 179–99.

Russo, Maria Chiara, Claudio Bendazzoli, Annalisa Sandrellim and Nicoletta Spinolo (2012) “The European Parliament Interpreting Corpus (EPIC): Implementation and Developments” in *Breaking Ground in Corpus-Based Interpreting Studies*, Francesco Straniero Sergio and Caterina Falbo (eds), Bern, Peter Lang: 35–90.

Sandrelli, Annalisa, Claudio Bendazzoli, and Mariachiara Russo (2010) “European Parliament Interpreting Corpus (EPIC): Methodological Issues and Preliminary Results on Lexical Patterns in Simultaneous Interpreting”, *International Journal of Translation* 22, no. 1–2: 165–203.

Schjoldager, Anne (1996) *Simultaneous Interpreting: Empirical Investigations into Target-Text/Source Text Relations*, Aarhus, Aarhus School of Business.

Seleskovitch, Danica, and Lederer Marianne (1989) *A Systematic Approach to Teaching Interpretation*, Paris, Didier.

Setton, Robin (1997) “A Relevance-theoretic Account of Simultaneous Interpretation”, *Tsuuyaku kenkyuu - Interpreting Research* 13, no. 2: 33–6.

---- (1999) *Simultaneous Interpretation: A Cognitive-pragmatic Analysis*, Amsterdam, John Benjamins.

Setton, Robin and Motta Manuela (2008) “Syntacrobatics: Quality and Reformulation in Simultaneous-with-Text”, *Interpreting* 9, no. 2: 199–230.

Shlesinger, Miriam (1989) “Monitoring the Courtroom Interpreter”, *Cahiers de l’Ecole de Traduction et d’Intérpretation* 11: 29–36.

---- (1998) “Corpus-based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies”, *Meta* 43, no. 4: 486–93.

---- (2008) “Towards a Definition of Interpretese. An Intermodal, Corpus-based Study”, in *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile*, Gyde Hansen, Andrew Chesterman, and Heidrun Gerzymisch-Arbogast (eds): 237–53.

Spinolo, Nicoletta (2014) *La resa del linguaggio figurato in interpretazione simultanea: Una sperimentazione didattica*, PhD diss., University of Bologna.

Straniero Sergio, Francesco, and Caterina Falbo (eds) (2012) *Breaking Ground in Corpus-based Interpreting Studies*, Bern, Peter Lang.

Timarová, Šárka (2005) “Corpus Linguistics Methods in Interpreting Research: A Case Study”, *The Interpreters' Newsletter* 13: 65–70.

Tonin, Raffaella (2010) *El vaivén de las palabras. Los anglicismos en español y en la traducción al italiano,* Roma, Aracne.

Viezzi, Maurizio (1993) “Written Translation and Simultaneous Interpretation Compared and Contrasted: A Case Study”, *The Interpreters' Newsletter* 5: 94–100.

Voncina, Katja (2009) *L'interpretazione simultanea al Parlamento europeo sull'esempio delle cabine tedesca, italiana e slovena*, MA diss., SSLMIT, Università degli Studi di Trieste.

Wadensjo, Cecilia (2001) “Approaching Interpreting through Discourse Analysis”, in *Getting Started in Interpreting Research. Methodological Reflections, Personal Accounts and Advice for Beginners*, Gile Daniel, Helle V. Dam, Friedel Dubslaff, Bodil Martinsen, and Anne Schjoldager (eds), Amsterdam, John Benjamins: 185–98.

Wallmach, Kim (2002) “Using Parallel Corpora to Determine Interpreting Strategies for Languages of Limited Diffusion in South Africa”, in *Proceedings of the Łódź Session of the 3rd International Maastricht-Łódź Duo Colloquium on “Translation and Meaning*, Barbara Lewandowska-Tomaszczyk, and Marcel Thelen (eds), Maastricht, Hogeschool Zuyd: 503–9.

Notes

[1] The corpus is currently being compiled by the author as part of a PhD project at the University of Bologna at Forlì, Dipartimento di Interpretazione e Traduzione (DIT). The corpus is built for a specific research purpose, though it falls within the context of the EPIC project (see footnote 4).

[2] By “unmodified English loanword” we make reference to Bombi (2005) and Furiassi's (2010) categorisation of anglicisms in Italian, where the lexical borrowing undergoes no modifications in the target language from the morphological and phonetic point of view. This kind of anglicism is often referred to as “*integrale*” since it is the most evident and the least adapted to the rules of the “importing” language.

[3] EPIC, the European Parliament Interpreting Corpus, is a trilingual (English-Spanish-Italian) machine-readable corpus developed at the University of Bologna at Forlì, under prof. Mariachiara Russo's supervision. It consists of online transcripts of original speeches delivered at the European Parliament and of the audio recordings of the related interpreted versions. The corpus is indexed, lemmatised and POS-tagged to make the retrieval of specific features easier and to make online consultation quicker (Sandrelli *et al.* 2010, Russo *et al*. 2012).

[4] For a detailed description of these transcription norms, see Bendazzoli (2010: 126).

[5] For each chart included in this paper, a frequency table is provided in the appendix.

[6] In some cases, the Italian speaker mispronounced the loanword completely, altering the British-American pronunciation (taken as a reference for “standard”) and even making it difficult for the recipient to recognize the anglicism as such.

[7] Weighted percentages were calculated by balancing the number of male/female speakers in the first case (Figure 8) and the number of speakers per political group in the second case (Figure 10): this way, the frequency of phenomena is not affected by the highest number of male speakers nor by the most represented political group and a balanced average of phenomena is provided for these two categories.

[8] Eurlex (**http://eur-lex.europa.eu/homepage.html**) is an online database available in 24 languages covering many types of texts produced mostly by the institutions of the European Union, but also by Member States, EFTA, and so on.

[9] Lexis Nexis is an online database of full-text documents from over 17,000 authoritative sources of information (mainly newspapers and press releases) in multiple languages from the early nineties to date (**https://www.lexisnexis.com/en-us/products/lexisnexis-academic.page** accessed 24/02/17).

[10] Real Academia Española. (2014). *Diccionario de la lengua española* (23.° ed.). **http://www.rae.es/rae.html** (accessed 22/02/17).

[11] Clave. (2014). *Diccionario de uso del español actual*. **http://www.smdiccionarios.com/home.php** (accessed 22/02/17).

[12] Real Academia Española y Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. **http://www.rae.es/obras-academicas/diccionarios/diccionario-panhispanico-de-dudas** (accessed 22/02/17).

"*ANGLINTRAD*: Towards a purpose specific interpreting corpus", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2317

"The TIPp project Developing technological resources based on the exploitation of oral corpora to improve court interpreting", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.
Stable URL: http://www.intralinea.org/specials/article/2316

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy.

The TIPp project

Developing technological resources based on the exploitation of oral corpora to improve court interpreting

By Mariana Orozco-Jutorán (Universitat Autònoma de Barcelona, Spain)

Abstract & Keywords

English:

In Spain, new laws have been passed that significantly reinforce procedural guarantees in criminal proceedings, as they provide regulation on the right to translation and interpreting in criminal proceedings as well as on the right to information of an accused person in relation to the subject of the criminal proceedings, so that they can exercise efficiently their right to self-defence. Translation and interpreting thus become an essential element in the right to effective legal protection in the exercise of lawful rights and interests before the courts in order to avoid any state of defencelessness.

In the light of this new situation, the research group MIRAS, of the Universitat Autònoma de Barcelona, launched a research project called TIPp (Translation and Interpreting in Criminal Proceedings) aimed at describing the reality of court interpreting and at creating a computer application which comprises all the necessary resources to facilitate court interpreters' performance. These include recommendations for court interpreters and courtroom personnel on interpreters' role and on how to interact with interpreters, monolingual glossaries in Spanish for different contexts, such as certain type of crimes or 'general vocabulary' for criminal trials and a pilot sample of five databases — one in each language combination (English, French, Romanian, Arabic and Chinese from and into Spanish) — containing the problematic units most frequently encountered by court interpreters, as observed in the TIPp corpus.

This article explains the design and the methodology used to compile and exploit the corpus, which will be made publicly available, as well as some of the results from the outcomes of this project.

Keywords: court interpreting, corpora, quality, detainee rights, ICT resources

Introduction

In Spain, court interpreting has been an under-researched area until recently. Academic contributions started barely a decade ago (Ortega Herráez 2006; Del Pozo Triviño et al. 2014; Onos 2014) and the descriptions of the current situation of court interpreting in Spain are not based on authentic, representative data. In the last two decades, however, research in court interpreting has emerged as a major topic in Europe. In fact, within the Horizon 2020 Programme, the Directorate-General for Justice of the European Commission, through the Justice Programme 2014-2020, is offering grants to undertake research in this area.

As a result of the transposition of two European Directives**[1]**, a new law was passed by the Spanish Parliament in April 2015 (*Ley Orgánica 5/2015, de 27 de abril*) amending Spain's Code of Criminal Procedure. As stated therein, this new legislation 'significantly reinforces procedural guarantees in criminal proceedings, as it provides regulation on the right to translation and interpreting in criminal proceedings as well as on the right to information of an accused person in relation to the subject of the criminal proceedings so that they can exercise efficiently their right to self-defence'**[2]**. Translation and interpreting thus become an essential element in the right to effective legal protection in the exercise of lawful rights and interests before the courts in order to avoid any state of defencelessness. This law is referring to the right to be informed of the accusation against a subject and the right to a public process with all procedural guarantees, as enshrined in Section 24 of the Spanish Constitution.

The research group MIRAS, of the Universitat Autònoma de Barcelona, is specialised in Public Service Interpreting, and the research projects previously undertaken by this group in Barcelona, Spain (for instance, see Onos 2014) revealed that court interpreters currently lack the required technological and research resources to carry out their tasks with accuracy, rigour and diligence. Furthermore, a review of the current literature, that is detailed in section 1, shows the absence of a description of reality using sufficiently representative data, that is, there are many assumptions and hypotheses being made about court interpreting but there is a lack of authentic and representative data to know what is actually happening.

1. The TIPp project

Given these needs, the research group MIRAS decided to launch a research project called TIPp**[3]** (Translation and Interpreting in Criminal Proceedings) aimed at compiling and analysing a representative oral corpus of trials in order to be able to describe the reality of court interpreting and at creating a computer application which comprises all the necessary resources to facilitate court interpreters' performance.

The researchers involved in the project had prior experience of projects using oral corpora from public service interpreting (see Arumí *et al*. 2011, 2012; Vargas-Urpi and Arumí Ribas 2014; Vargas-Urpi 2012) but TIPp is the first project based on collection and use of authentic oral corpora in court interpreting settings. There are four features of the TIPp project that make it unique.

The first is the novelty of being able to access real, video-recorded criminal proceedings. This is a breakthrough for court interpreting research, and has required a great deal of effort because, as Angermeyer, Meyer and Schmidt (2012: 276) point out:

> Permissions for tape-recording sensitive data from medical or juridical communication can usually be obtained only after long, strenuous negotiations with the respective institutional bodies, and it surely can be assumed that many research projects have turned out to be not feasible simply because of bureaucratic hindrances.

The second feature is the size and representativeness of the oral corpus. The literature available (Berk-Seligson 1987, 1988, 1989 and 1999; Cooke 1995; Goldflam 1995; Hale 1997a, 1997b, 1997c, 1999, 2002 and 2008; Kadric 1999; Lane, McKenzie-Bridle and Curtis 1999; Mikkelson 1998; Montalvo 2001; Morris 1999; Nicholson and Martinsen 1997; Niska 1995; Ortega 2006 and 2011; Rigney 1999; Stern 1995, to name a few) shows that the studies on court interpreting based on oral corpora conducted so far have yielded very interesting and insightful data, such as the role usually played by the interpreter in a courtroom, that goes from mere conductor to "assistant" of the courtroom personal or even "mediator". However, they have mostly been based on corpora that are either simulated – and thus cannot be claimed to describe reality – or relatively small – and thus cannot be used to extrapolate results or claim significance from the point of view of research methodology.

There are only two known exceptions to this, the first is a study conducted by Berk-Seligson (1990/2002), where the author investigated how Spanish-English interpreters faced the challenges of legal discourse in 114 hours of interaction in US courtrooms, highlighting the influence on the receivers' perceptions of the way in which people spoke and were interpreted. The second exception is Angermeyer's study (Angermeyer 2015: 6), where the researcher observed over 200 court proceedings and tape-recorded 60 hearings and transcribed them. The main difference between this study and TIPp is that Angermeyer observed

small claims courts, so the cases studied were mostly arbitration hearings, whereas the TIPp corpus is based on criminal proceedings in criminal courts.

Since one of the TIPp project's declared aims is to describe reality using representative data, researchers chose to create and exploit a significant, representative corpus of real criminal proceedings that had very recently taken place. TIPp has accessed the video-recordings of criminal trials where interpreting took place in almost half of the criminal courts in Barcelona from 2010 to 2015. The corpus is described in depth below, in section 3.

Due to the importance and the difficulty of having access to a representative oral corpus of real criminal proceedings, the corpus compiled, transcribed and annotated will be made available for researchers so that it can be used in the future.

The third feature of the project consists of the systems used for the transcription and annotation of the corpus. In order to obtain quantifiable data to be able to describe reality in a systematic rather than an anecdotic way, the research team chose to use one tool that not only facilitated the transcription and annotation of the corpus but also allowed the creation of ad hoc categories for the annotations. This all-inclusive tool is a software package called EXMARaLDA, a system for the computer-assisted creation and analysis of spoken language corpora**[4]**. This tool enables the user to compile and manage a corpus, transcribe videos and, most importantly, it facilitates the type of ad hoc annotation created as well as its conversion into quantifiable data. Details of the transcription and annotation are explained below, in sections 4 and 5.

Finally, the fourth feature is the number of resources created. As well as describing reality, the project aims to provide support for users of interpreting in court settings by creating a computer application which includes resources to improve court interpreters' performance. These resources include (i) a set of recommendations for court interpreters, (ii) a set of recommendations for courtroom personnel regarding the role of the interpreter and how to interact with interpreters, (iii) monolingual glossaries in Spanish for different contexts, such as certain type of crimes or "general vocabulary" for criminal trials, for instance and (iv) a pilot sample of five databases -one in each language combination (English, French, Romanian, Arabic and Chinese from and into Spanish)- containing the problematic units most frequently encountered by court interpreters, as observed in the TIPp corpus. This freely accessible resource is described below, in section 6.

2. Corpus compilation

After a long process of interaction with the judicial institutions involved, researchers were able to request access to the video-recordings of criminal trials where interpreting took place in criminal courts in Barcelona. Criminal proceedings have been video-recorded in Barcelona's criminal courts since 2009 and these recordings constitute the official records of the proceedings. Obtaining permission to access the video recordings involved several meetings and submission of written documents explaining very clearly the interests of the researchers and the use that would be made of the oral corpus to be created, as well as the commitment to anonymise the corpus by signing a strict confidentiality agreement. Attention was focused on a specific criminal summary procedure known in Spanish as *procedimiento abreviado* and specifically the cases tried in courts known as *Tribunales de lo Penal* (Criminal Courts).

Once permission to access video-recordings was granted, the researchers first studied the listings of all the trials that had taken place in the last seven years (2009-2015) as provided by the service of translation and interpreting of the Justice Department. They then selected those in which interpreting in the five working language combinations of the research team (Romanian, English, Arabic, Chinese and French) had supposedly taken place. Finally, the decision was made to request the recordings available from 50% of the *Tribunales de lo Penal* in Barcelona, so that the corpus compiled would be representative. There are currently 28 such courts in Barcelona, but only 24 of them are specifically trial courts, since four of them are devoted to enforcement of judicial resolutions which are solved through written proceedings. Therefore, out of a total of 24 such courts where interpreting is used, the researchers requested the video-recordings of 12 criminal courts, which were chosen randomly.

2.1. Unexpected events

In principle, the methodology for the corpus design thus satisfied all research requirements for representative data collection in time. However, unexpected events modified this initial situation, and although researchers can still claim to have a representative corpus that can describe reality, because it includes 50% of all data available, the size of the corpus was considerably diminished as a result of the following circumstances.

Firstly, although permission had been granted for access to video-recordings of the 12 courts chosen randomly, each court had to be provided with a specific list of the videos requested. Each court had its own clerks and its own method of dealing with the working processes for which they were responsible. Consequently, some of the courts were very quick to provide the videos requested whilst in others the process took several months. In fact, two of the selected courts were finally eliminated from the list because, after waiting for over one year, they were unable to deliver the recordings requested, due to administrative and bureaucratic problems.

Secondly, after receiving the video recordings from each selected court and checking them against the list of recordings that had been requested, the researchers found that some videos were missing, so that each remaining court was sent a second request for the missing videos. Finally, only after a further year was it possible to complete an electronic folder including all the videos received.

Thirdly, given the late reception of the video-recordings, the process of transcription was delayed until June 2015. Moreover, when researchers began studying the videos, they discovered that quality of sound and image of many of them, especially the oldest ones, made it very difficult to transcribe them and to create a corpus that could then be used for the proposed description of reality. The decision was thus made to use only the most recent recordings (2014-2015) which were of a better quality.

A further problem related to the transcription phase arose when the large number of video-recordings received was checked against the funds available to pay technicians to transcribe the recordings. This led to the difficult decision to start by transcribing the 2015 videos and to transcribe only three language combinations (English, French and Romanian) instead of the five initially conceived**[5]**.

3. Corpus description

Although it was not possible to transcribe many of the video-recordings obtained because of lack of time and funds, there was nevertheless some very interesting metadata that could be obtained from them. Therefore, a list of 20 items was created and the TIPp project, besides transcribing the 2015 trials, is also extracting this metadata from all the video recordings received, from 2010 to 2015. The metadata includes items such as the quality of sound and image, the interpreting techniques used (*chuchotage*, notetaking) if the interpreter is introduced by the judge or not, and other data that could be of interest for further studies or for deciding which trials to transcribe in possible future projects.

The transcribed corpus, thus, includes the first six months' recordings from 2015. This, however, does not mean that more trials will not be transcribed in the future, if more funds are found to enlarge the corpus.

Therefore, the final, transcribed corpus includes all the videos obtained from trials where interpreting took place in 10 criminal courts of Barcelona for three language combinations (English, French and Romanian into Spanish). Table 1 illustrates the characteristics of the corpus that has been actually transcribed. The researchers hope that this work in progress can evolve in the future to include the transcription of recordings in the other two language combinations (Arabic and Chinese into Spanish) in the corpus.

	A	B	C	D	E	F	G	H
2015	Trials where an	Missing	Video	Trials	with	Trials	Transcribed	Bilingual	Total

(January-June)	interpreter was requested	video recordings	recordings obtained	no actual interpreting	with interpreter	Trials	minutes transcribed	minutes of trial transcribed
French	52	9	43	32	11	9	92	190
English	65	10	55	33	22	19	123	371
Romanian	114	37	77	45	32	27	124	555
TOTAL	231	56	175	110	65	55	335	1116

Table 1. Transcribed 2015 corpus description

The first column of table 1 shows the period in which the trials were video-recorded and the linguistic interpreting combination, in all cases into and from Spanish. Column A shows the number of hearings in which, according to the listings provided by the service of translation and interpreting of the Justice Department, an interpreter was requested (a total of 231). However, when recordings were requested, even the second time, many were missing, and column B shows the number of final missing recordings (a total of 56). We cannot be absolutely sure about the reasons for these missing video recordings, but it is very likely that this may be due to trials that were suspended, for example, because the defendant or his/her lawyer did not show up. The result of subtracting the missing recordings (column B) from the original list of potential recordings (column A) is column C, which shows the actual number of videos obtained. These vary from 43 in the French-Spanish language combination to 77 in the Romanian-Spanish language combination, a total of 175 trials.

A further number has to be subtracted (column D) from these 175 recorded trials because the researchers noticed that in effect there was no intervention of an interpreter. The reasons in this case are known and vary from cases where the witness who was going to be interpreted did not appear in court, to cases where a plea bargain agreement was reached between the parties before the trial started and therefore the intervention of the interpreter was unnecessary. The resulting number is surprising since it represents almost two thirds of the trials, so that once subtracted from the initial number, only 65 include the intervention of an interpreter (column E). The marked difference between the official data made available in the list of trials in which an interpreter was requested (231) and the trials in which an interpreter was actually involved (65) should be taken into account when describing reality. This article does not aim to discuss the results of the data obtained, but we believe they are worthy of note, since they have clear implications that will be discussed in further articles.

Finally, of the 65 recordings in which an interpreter was involved, some were not transcribed because either the interpreter did not have to speak -because the accused or the witness said that s/he could speak in Spanish and did not need an interpreter- or the only interpreting taking place during the trial was *chuchotage*, which is not recorded in the video because the volume is too low to be recorded. Ultimately, therefore, the corpus gathered consists of the transcription of 55 trials (column F) which altogether last for 1116 minutes (column H).

Column G shows the difference between the total duration of the trials (that amount to 1116 minutes of oral interventions that have been transcribed) and the total minutes interpreted, which only amount to 339. If we add the number of minutes where there has been *chuchotage*, this figure grows to 513 minutes, which is 46% of the total minutes of the trials. This data is of interest because it means that less than one halve of the trial is actually interpreted to the defendant, a finding that implies a clear violation of the defendant's right of information according to both European and Spanish laws.

In sum, the TIPp transcribed corpus consists of 55 trials and 1116 minutes of oral interventions in three language combinations: English, French and Romanian from and into Spanish.

Given the amount of material gathered and the impossibility of transcribing all the 2014 video-recordings, the researchers decided to leave the transcription of the 2014 videos for further research projects. Nevertheless, the main data in these videos, as displayed in Table 2, was extracted and proved to be very helpful when determining whether the 2015 corpus was really representative and significant in terms of the description of reality. A comparison was made between the number of trials that finally did not take place for whatever reason in 2014 and 2015 as well as the numbers of trials in which an interpreter was requested but finally was not needed.

2014 (January-December)	Trials where an interpreter was requested	Missing video recordings	Video recordings obtained	Trials with no interpreting	Trials with interpreter
Arabic	258	97	161	77	84
Chinese	97	36	61	17	44
French	75	44	31	18	13
English	77	19	58	31	27
Romanian	206	89	117	68	49
TOTAL	**713**	**285**	**428**	**211**	**217**

Table 2. Not transcribed 2014 corpus description

Table 2 shows that the data obtained for 2014 supports the reliability of the data obtained in 2015, since the proportion between the total figures for both items is very similar.

4. Corpus transcription

There are several transcription systems available to researchers, and the differences between them can be very small regarding, for instance, how to represent a pause or a fragment that cannot be understood inside the transcription, but there can also be important theoretical variations**[6]**.

Researchers in the TIPp project, after considering the different possibilities, decided to transcribe in the simplest and most straightforward way possible, since the main interest was in the annotation of a corpus that reflected reality. This means writing what is said exactly as it is heard. Thus, for instance, grammar mistakes, incorrect pronunciation or hesitations are transcribed without amendments or comments. There is only one exception to this rule, when incorrect pronunciation causes the reader of the transcription to misunderstand what is being said. For example, in one case in which the accused says what sounds like "aquachis" meaning "aquagym", the transcriber has to write "aquagym" but also to include the word as it was pronounced between square brackets. When a word is incomprehensible, due to problems of the sound recording –for instance during *chuchotage* if the interpreter talked at a distance from the microphone- then the transcriber marks it with three points inside round brackets: (…).

Another important decision researchers made was not to include any reference to nonverbal communication, unless it was completely necessary in order to understand the message. An example of this necessary comment would be the case of the accused shaking his head to say "no" but not saying anything and the interpreter saying "no". In this case, the transcriber

includes a comment explaining in a very simple way what has happened, between double round brackets, for example ((the accused moves his head from right to left meaning "no")).

Finally, the most important difference between the chosen transcription system and other possible options is that all the TIPp corpus is fully anonymised, for confidentiality reasons. This is also important in order to be able to make the corpus publicly available. Therefore, all references to names of people, streets, recognisable places such as restaurants, badge numbers of policemen, telephone numbers and so on have been substituted by a list of fake, previously accorded names and numbers.

Regarding the software used, only one tool was used to transcribe, annotate and retrieve the data desired: EXMARaLDA**[7]**. EXMARaLDA was originally developed in the project "Computer-assisted methods for the creation and analysis of multilingual data" at the Collaborative Research Center "Multilingualism" (Sonderforschungsbereich "Mehrsprachigkeit" – SFB 538) at the University of Hamburg and since 2011, the development of EXMARaLDA continues at the Hamburg Centre for Language Corpora in cooperation with the Archive for Spoken German at the Institute for the German Language in Mannheim. It consists of a transcription and annotation tool (Partitur-Editor), a tool for managing corpora (Corpus-Manager) and a query and analysis tool (EXAKT). It works with XML based data formats which interoperate with one another and enables a flexible processing and sustainable usage of the data. This was a major finding for the TIPp project, since the tool allows researchers to create, annotate and exploit the corpus at one and the same time, and even facilitates the extraction of specialized terminology, which is also important for one of the outputs of the TIPp project: the creation of terminological records.

Figure 1 shows an example of a fragment of a trial transcribed, using the EXMARaLDA software.

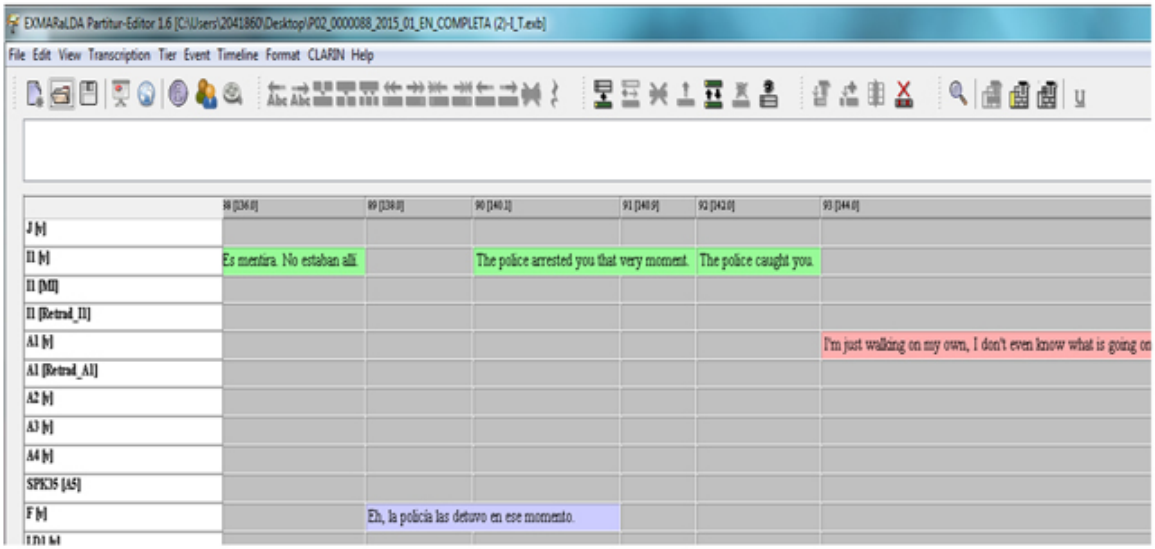


Figure 1. Fragment of a trial transcription using the EXMARaLDA software.

The example transcribed in Figure 1 shows how a different colour has been assigned to every speaker, so the green colour refers to the interpreter, who says 'Es mentira. No estaba allí' [*That is not true, I wasn't there*], then the prosecutor, in blue colour, says 'Eh, la policía las detuvo en ese momento' [*Eh, the police arrested them at that moment*] and then the interpreter, before the prosecutor finishes the sentence, starts speaking again to translate what the prosecutor just said. The overlap between the speakers can be seen thanks to the timeline provided by the EXMARaLDA software. Then, the accused person, in red colour, says 'I'm just walking on my own, I don't even know what's going on'.

As can be also seen in Figure 1, there is one tier or row devoted to each speaker, so that all trials consulted can be easily analysed, because they always follow the same order: the first tier or row is devoted to the Judge, the second to the interpreter, the third to a second interpreter in case there are two interpreters in the room, the fourth to the retranslation of the interpreter into Spanish (this is only used when needed, for instance in the case of Romanian, Chinese and Arabic, which are languages not so well known for all the researchers, but it is not used when the language to which the interpreter is translating is English or French), the fifth to the accused, the sixth and the seventh to other possible accused people, the eighth to the prosecutor, the ninth to the defence lawyer, and so on.

5. Corpus annotation

Regarding annotation, the researchers first checked many different annotation systems, such as part-of-speech, lemmatization, syntactical (parsing), semantic (domain classifications), coreference (discourse), pragmatic (speech acts – dialogue) and stylistic**[8]**, and then also considered other qualitative content analysis annotation systems, but found that, although some of the latter systems were close to the needs of the TIPp project, none was suitable for the study purposes.

Therefore, an ad hoc annotation system for this research was created from zero. The main goal of the project of describing reality was operationalised into categories or indicators that can be observed and marked in the corpus, and a whole classification system was created. This system includes, first of all, two main categories, namely interaction and textual problems, based on Wadensjö's distinction between 'talk-as-activity' and 'talk-as-text' (Wadensjö 1998: 21).

The textual problems annotated assess the fidelity of the message conveyed by the interpreter and signal the places in which the interpreter has found linguistic, cultural, or domain-related (for instance legal) problems in the oral discourse. Here, linguistic is understood in the wider sense of the term, including not only textual, syntactic and lexical levels, but also the pragmatic level, so that it would include, for example, problems of register or changes in the discourse.

The textual problems are firstly tagged and then two different annotations are marked and stored in the corpus for each element; the first one, shown in Table 3, assesses if the solution to the textual problem found has been (i) adequate, that is, conveying the message adequately, (ii) inadequate, that is, not conveying the message adequately or (iii) improvable, that is, the interpreter conveys the message roughly but the solution could be clearly improved.

Textual annotation:
1. Indicator of fidelity, that is the solution applied by the interpreter when facing a textual problem was:
- (A) Adequate.
- (M) Improvable
- (I) Inadequate.

Table 3. Scale created and used to annotate in the corpus the solution applied by the interpreter when facing a textual problem.

The second annotation for textual problems signals the type of solution adopted by the interpreter and the possible categories are shown in Table 4.

Textual annotation: 2. Indicator of the type of solution applied by the interpreter when facing a textual problem:
Possible categories when the solution applied has been marked in the previous textual indicator as 'adequate': - (EH) Usual equivalent. - (IM) Making some information implicit. - (EX) Making some information explicit.
Possible categories when the solution applied has been marked in the previous textual indicator as 'improvable': - (CR) Change of register - (NMS) Slightly different meaning (from that of the original message).
Possible categories when the solution applied has been marked in the previous textual indicator as 'Inadequate': - (O) Omission. - (OG) Dangerous omission. - (NT) Not translated. - (AD) Addition of information. - (ADG) Dangerous addition of information. - (ITER) Inadequate terminology. - (FS) Wrong meaning (a very different meaning from that of the original message). - (FSG) Dangerous wrong meaning. - (CS) Opposite meaning (saying the opposite of what was conveyed in the original message). - (SS) Sentence with no meaning (message is not understandable, does not make sense).

Table 4. Scale created and used to annotate in the corpus the type of textual solution applied by the interpreter when facing a textual problem.

As shown in Table 4, there are many possible solution types that have been annotated and stored in the corpus. Unfortunately, we cannot describe them thoroughly here, since a whole article is needed to do that, so in order to see a thorough explanation of these categories and examples of each of them see Orozco-Jutorán (2017b).

However, we would like to point out that there has been a distinction made between 'serious errors' (which include four of the categories listed in Table 4 as inadequate types of solutions: dangerous addition of information, dangerous omission, sentence with no meaning and dangerous wrong meaning) and other, 'less serious' type of errors. By serious errors, we mean errors that might affect or interfere with the result of the proceeding, as shown in the following example, where we have included our translation of the Spanish oral interventions between square brackets and where the dangerous addition of information is underlined:

Judge: *... que si reconoce los hechos y está conforme.*
[Does he acknowledge the facts and agrees?]
Interpreter: *Do you accept?*
Defendant: *Sí.*
[Yes]
Interpreter: *Yeah? And do you agree?*
Defendant: *Yeah.*
Interpreter: *Sí, es culpable.* [Yes, he is guilty]

Although there is no space here to make an analysis of the results found, we would like to signal that the amount of serious errors found in the corpus is alarmingly large, as Table 5 shows.

Language	Dangerous omissions per bilingual hour	Dangerous addition of information per bilingual hour	Dangerous wrong meanings per bilingual hour	Sentences with no meaning (SS) per bilingual hour	**Total of serious errors per bilingual hour**
English	6,3	2,6	7,3	4,4	20,6
French	5,9	1,3	6,5	1,3	15,0
Romanian	12,6	4,8	7,3	1,0	25,7
Mean	**8,5**	**3,2**	**7,1**	**2,3**	**21,1**

Table 5. Number of serious errors found in the corpus per bilingual hour of trial.

We would also like to mention that one of the findings yielded by the analysis of the annotated corpus is that most of the trial is not actually translated for the user, who is usually the defendant. This is measured by one of the categories created under "inadequate textual solutions": "not translated" (NT). In order to be marked as NT, there needs to be a whole intervention (therefore, a whole speech act) by the judge, the defence lawyer, the public prosecutor or a witness which has not been translated, so there is an important difference with the omissions, which affect only a word or a sentence which has not been translated. Table 6 shows the amount of NT found per hour and per minute in the corpus, which, again, is alarmingly large.

Language	**Total of NT per hour**	**Total of NT per minute**
English	371	1,8
French	190	1,6

Romanian	555	3,7
Mean	**372**	**2,7**

Table 6. Number of "Not translated" interventions (NT) found in the corpus per hour and per minute of trial.

The interaction problems annotated signal the moments in the oral interaction in court where any one of the participants (judge, lawyers, interpreter, defendant, witnesses, and so on) has had a problem. These problems include those relating to conversation management, non-renditions (Wadensjö 1998) and speech style. In order to annotate each of these types of problems, several categories were created, as Table 7 shows.

Interaction annotation:
Possible categories regarding conversation management problems: - (S) overlap - (I) Interruption - (DL) long turns (that is when a member of the judicial staff speaks for more than two minutes in a single turn)
Possible categories regarding conversation non renditions: - (J) Justified (that is pause, clarification, confirmation or retrieval) - (NJ) Unjustified (that is warning, instructions, advice, answering on behalf of the defendant or adding extra information) - (RT) Reactive tokens (that is when the interpreter's non-rendition merely acknowledges that he or she received the information in the original utterance)
Possible categories regarding speech style, by both the interpreter and the courtroom personnel: - Direct speech - Indirect speech - Reported speech

Table 7. Scale created and used to annotate interaction problems in the corpus.

Again, we cannot describe the categories thoroughly here, since a whole article is needed to do that, so in order to see a thorough explanation of these categories and examples of each of them, see Arumí and Vargas-Urpí (forthcoming).

Figure 3 shows what the annotations look like in the corpus. As can be seen, one tier or row is devoted to each of the types of problems mentioned, both textual and interaction problems. In the example, on top of Figure 3, there are all the tiers or rows devoted to the speakers and the transcription of what they said at the fragment previously shown in Figure 1. Then, below those rows, starting in tier 17, the annotation tiers can be seen, the first one called 'PROBLEMA'. This tier is where the researchers tag the fragment where there is a textual or interaction problem. For instance, on the first grey column in Figure 3, below where the interpreter says 'Es mentira. No estaba allí', there is an 'I' meaning that there is an 'interaction problem' in that sentence, and then, a few rows below, in the tier devoted to speech style, there is the annotation *INDIR*, meaning that the interpreter is using indirect speech (saying '*They* were not there' instead of using the same speech style used in the original sentence by the defendant, which would be '*We* were not there').

In the next column, to the right, the prosecutor speaks, saying 'Eh, la policía las detuvo en ese momento' [*Eh, the police arrested them at that moment*] with no annotation or tag below, because there is nothing to be annotated in that sentence, since there is not any problem faced by the interpreter there, and then in the next column, the interpreter translates the prosecutor but starts speaking before the prosecutor finished his sentence. This overlap between the speakers is marked by the tag "I" in the tier '*PROBLEMA*', since there is an interaction problem, and then there is the tag *SOI* at the tier belonging to *SOLAPAMIENTO*, which means "overlap" in Spanish. This *SOI* stands for "overlap with the interpreter", and is differentiated from an overlap between the Judge and the prosecutor of the defence attorney, which would be annotated as *SOJ*. In this same sentence there are two more annotations. The first one is not an interaction problem but an interaction observable phenomenon (and that is why, in the tier for *PROBLEMA*, next to the "I", there is an "F", which stands for *Fenómeno*, which is the Spanish word for "phenomenon"). The observable phenomenon is then annotated in the style tier, tagged as *DIR*, because here the interpreter is not using indirect speech but direct speech, as would be recommended in this case. The second annotation is of textual nature, that is why next the "I" and the "F" at the *PROBLEMA* tier there is also an "S" (meaning "Solution"). In the tier right below this one, there is the annotation "A", meaning "adequate solution" and in the tier below the type of solution applied by the interpreter to the textual problem is tagged as EH, which stands for "Equivalente Habitual" [*usual equivalent*].

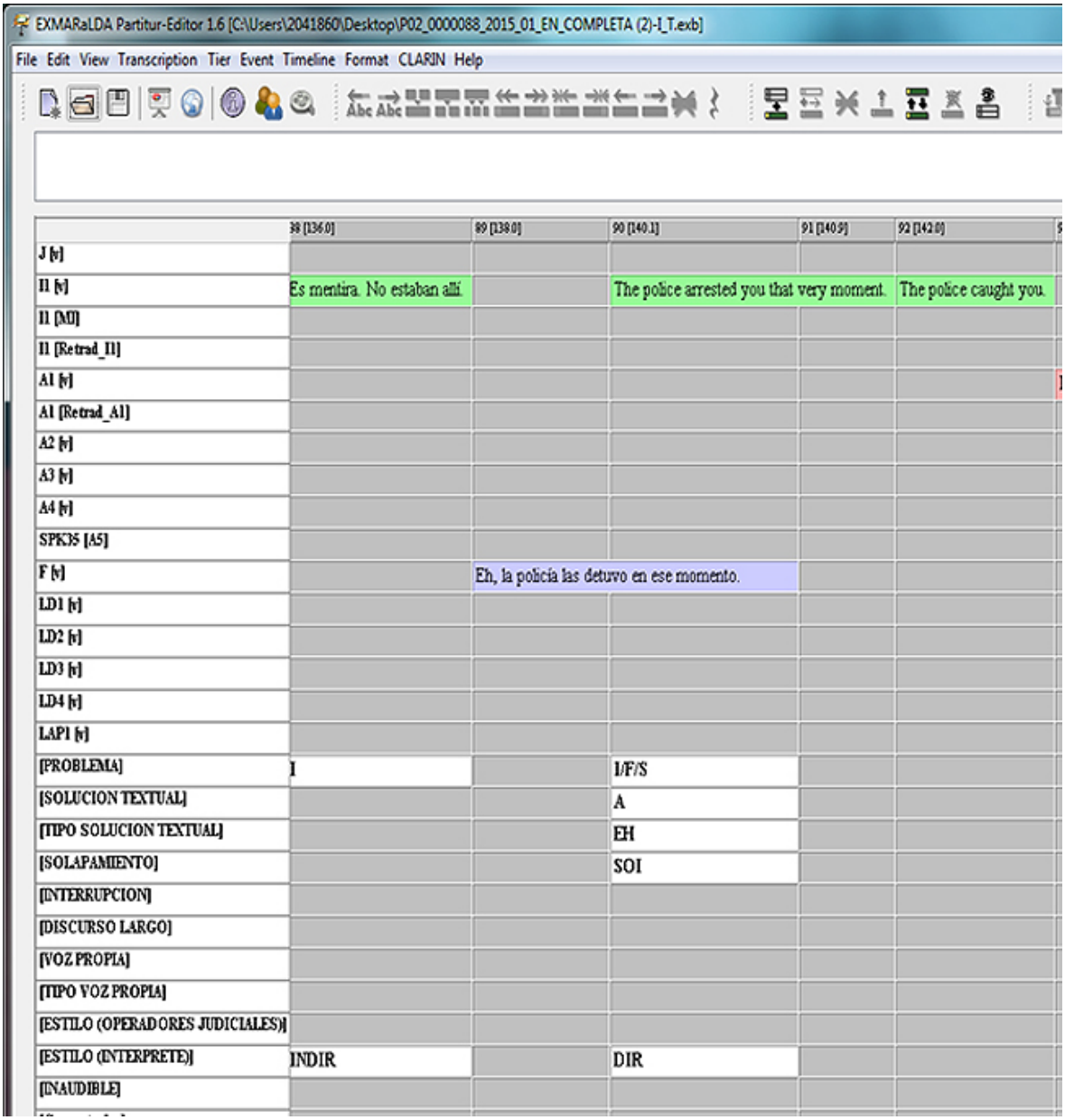


Figure 3. Fragment of a trial transcribed and annotated.

All the information annotated in the corpus in the way that has just been explained is then converted or transformed into excel files, an example of which can be seen at Figure 3.

	A	B	C	D	E	F	G	H	
1	PROBLEMA	SOLUCION_	TIPO_SOLUCION	SOLAPAMIEN	INTER	DISCURS	VOZ_PROPIA	TIPO_VOZ_PROPIA	ESTILO_OPERAD
2	F								DIR
3	I								INDIR (FP)
4	I								
5	I								INDIR (FP)
6	F								
7	I								INDIR (FP)
8	I/T	I	NT				NJ	INFOEXTRA	
9	T	I	NT						
10	F						J	ACLAR	
11	I				IR				

Figure 3. Detail of an excel file that includes the annotations.

There is an excel sheet for each trial and then one excel book or file containing all the trials in one language combination. Then, there is a "bigger" excel file, linked to the three sheets containing total data for each language, that combines the results of the three language pairs that have been analysed. As can be seen in Figure 3, the rows or tiers from the EXMARaLDA software are converted here in columns and allow the application of filters and formulas to obtain quantifiable data as the one shown in tables 5 and 6. This system has proved to be very useful because it allows researchers to obtain quantifiable data to be able to describe reality in a systematic rather than an anecdotic way.

6. Resources

As has been already mentioned, the TIPp project aims at describing reality but also wishes to contribute to improving court interpreters' performance by creating a series of resources directed to interpreters and courtroom personnel. TIPp has created a free, accessible website designed to be used from any mobile device that includes four resources.

Firstly, a set of recommendations for court interpreters, which could be considered a code of good practice, that is, a protocol for professional conduct in the most frequent situations for a court interpreter. The difference between the already existing codes and this resource is that TIPp's intention is to give focused, practical advice that can be applied by interpreters in their daily performances and that all advice given is based on irregular or difficult situations observed in the corpus compiled and analysed, and therefore respond to real court interpreters' needs. The recommendations are specific, written suggestions or videos.

Secondly, the same procedure has been followed to write a set of recommendations for courtroom personnel regarding the role of the interpreter and how to interact with interpreters.

Thirdly, the website contains Spanish monolingual glossaries for different contexts, such as certain type of crimes or "general vocabulary" for criminal trials for example. Each of these glossaries includes lists of terms found in the corpus and examples of use for each term, so that collocations and context can be seen by the interpreter in order to help him/her when preparing for court interpreting.

Lastly, the application includes a pilot sample of five databases -one in each language combination (English, French, Romanian, Arabic and Chinese from and into Spanish)- containing the problematic units most frequently encountered by court interpreters, as observed in the TIPp corpus. The databases include, for every term or unit, a translation-oriented terminological record which includes potential solutions, comments and translation options, following the structure of the translation-oriented record created for a previous research project**[9]** (for further explanations on this type of record, see Prieto and Orozco-Jutorán 2015 and Orozco-Jutorán 2017a). Although researchers initially intended to create exhaustive databases in five language combinations, the decisions made by researchers (explained in sections 2 and 3) have meant that the current corpus is only exhaustive for three language pairs (English, French and Romanian into Spanish). Therefore, the terms included for Chinese and Arabic are only a small sample, taken from transcription of case studies. The researchers aim to enhance the databases by adding more transcriptions to the corpus in the future, provided more funds are made available for transcribing a larger number of trials.

5. Conclusions

The TIPp project has used a pioneering methodology in the field of court interpreting research in Spain, as it is based on authentic materials extracted from real criminal proceedings. These materials have allowed the researchers to create and exploit a representative oral corpus that can be further extended in the future. On the basis of what has been observed in the corpus, through the use of an ad hoc annotation system that marks both textual and interaction problems found by the court interpreters, the researchers have reached important and alarming conclusions, such as that less than half of the hearing is actually interpreted to the defendant, that only 30% of the interpretation is audible and is properly recorded, or that there are too many serious errors in the translated part of the trials for it to be considered acceptable, which actually means that there is a violation of the defendant's rights.

Besides this descriptive data, the researchers have used the information obtained from the corpus to create a computer application, accessible through any mobile device, that includes resources directed to both interpreters and courtroom personnel which aims at helping court interpreters to perform their tasks more accurately and efficiently.

We hope that this will subsequently have an impact on the main users, namely defendants, usually from migrant communities, who could be left defenceless unless provided with effective legal protection through the services of good quality, professional interpreting in the courtroom setting. Furthermore, other secondary users of interpreting during trials, such as witnesses and victims will also benefit from the improved quality of court interpreting.

References

Angermeyer, Philipp S. (2015) *Speak English or What? Codeswitching and Interpreter Use in New York City Courts,* New York, Oxford University Press.

Angermeyer, Philip S., Bernd Meyer, and Thomas Schmidt (2012) "Sharing Community Interpreting Corpora. A Pilot Study" in *Multilingual Corpora and Multilingual Corpus Analysis*, Thomas Schmidt and Kai Wörner (eds), Amsterdam, John Benjamins: 275–294.

Arumí, Marta, Carmen Bestué, Sofía García-Beyaert, Anna Gil-Bardají, Jacqueline Minett, Liudmila Onos, Begoña Ruiz de Infante, Xus Ugarte, and Mireia Vargas-Urpi (2011) *Comunicar en la diversitat. Intèrprets, traductors i mediadors als serveis públics*, Barcelona, Linguamón-Casa de les Llengües. URL: **http://grupsderecerca.uab.cat/miras/sites/grupsderecerca.uab.cat.miras/files/informe_miras_ispc_2011_0.pdf** (accessed 10 June 2016).

Arumí, Marta, Carmen Bestué, Sofía García-Beyaert, Anna Gil-Bardají, Jacqueline Minett, Miren Olaciregui, Liudmila Onos, Begoña Ruiz de Infante, Xus Ugarte, and Mireia Vargas-Urpi (2012) "Traducció i immigració: La formació de traductors i intèrprets als serveis públics, noves solucions per a noves realitats" in *Recerca i immigració IV.* Barcelona, Generalitat de Catalunya: 157–183.

Arumí, Mireia and Marta Vargas-Urpi (forthcoming) "Annotation of Interpreters' Conversation Management Problems and Strategies in a Corpus of Criminal Trials in Spain: The Case of Non-renditions", *Translation and Interpreting Studies* 13, no. 3.

Bendazzoli, Claudio (2010) *Corpora e interpretazione simultanea*, Bologna, Asterisco. URL: **http://amsacta.unibo.it/2897/** (accessed 10 June 2016).

Berk-Seligson, Susan (1987) "The Intersection of Testimony Styles in Interpreted Judicial Proceedings: Pragmatic Alterations in Spanish Testimony", *Linguistics* 25: 1087–1125.

---- (1988) "The Impact of Politeness in Witness Testimony: The Influence of the Court Interpreter", *Multilingua* 7, no. 4: 411–439.

---- (1989) "The Role of Register in the Bilingual Courtroom: Evaluative Reactions to Interpreted Testimony" in *U.S. Spanish: The Language of Latinos. Special issue of the International Journal of the Sociology of Language*, Irene Wherritt and Ofelia Garcia (eds) 79, no. 5: 79–91.

---- (1999) "The Impact of Court Interpreting on the Coerciveness of Leading Questions", *Forensic Linguistics* 6, no.1: 30–56.

---- (1990/2002) *The Bilingual Courtroom: Court Interpreters in the Judicial Process*, Chicago and London, University of Chicago Press.

Cooke, Michael (1995) "Interpreting in a Cross-cultural Cross-examination: An Aboriginal Case Study", *International Journal of the Sociology of Language* 113, no. 1: 99–111.

Del Pozo Triviño, Maribel, Antonio Vaamonde List, David Casado-Neira, Silvia Pérez Freire, Alba Vaamonde Paniagua, Doris Fernandes del Pozo, and Rut Guinarte Mencía (2014) *Communication Between Professionals Providing Attention and Gender Violence Victims/Survivors Who Do Not Speak the Language: A Report on the Survey Carried Out on Agents During the Speak Out for Support (SOS-VICS) Project.* Vigo: Servizo de Publicacións da Universidade de Vigo.

Edwards, A. Jane, and Martin D. Lampert (eds) (1993) *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale NJ, Lawrence Erlbaum Associates.

Edwards, A. Jane (2001) “The Transcription of Discourse”, in *The Handbook of Discourse Analysis,* Deborah Schiffrin, Deborah Tannen and Heidi E. Hamilton (eds), Malden MA, Blackwell: 321–348.

Emerson Crooker, Constance (1996) *The Art of Legal Interpretation. A guide for court interpreters*, Portland OR, Portland State University.

Fairclough, Norman (2001). “Critical Discourse Analysis”, in *How to Analyse Talk in Institutional Settings: A Casebook of Methods*, Alec McHoul and Mark Rapley (eds), London, Continuum: 25–40.

Falbo, Caterina (2005) “La transcription: une tâche paradoxale”, *The Interpreters' Newsletter* 13: 25–38.

Fowler, Yvonne (2007) “Interpreting into the Ether: Interpreting for Prison/Court Video Link Hearings”. Presentation at the *Critical Link 5 – Quality in Interpreting: A Shared Responsibility*, 11-15 April 2007 Parramatta – Sydney (Australia). URL: **http://static1.squarespace.com/static/52d566cbe4b0002632d34367/t/5347f7e7e4b0b891fcd56cee/1397225447306/CL5Ellam_Fowler.pd** (accessed 10 June 2016)

Goldflam, Russell (1995) “Silence in Court! Problems and Prospects in Aboriginal Legal Interpreting” in *Language in Evidence: Issues confronting Aboriginal and multicultural Australia*, Diana Eades (ed), Sydney, University of New South Wales Press: 28–54.

Hale, Sandra (1997a) “Interpreting Politeness in Court. A Study of Spanish-English Interpreted Proceedings” in *Research, Training and Practice. Proceedings of the 2nd Annual Macarthur Interpreting and Translation Conference*, Stuart Campbell and Sandra Hale (eds), Milperra, UWS Macarthur/LARC.

---- (1997b) “Clash of World Perspectives: The Discursive Practices of the Law, the Witness and the Interpreter”, *Forensic Linguistics* 4, no. 2: 197–209.

---- (1997c) “The Treatment of Register Variation in Court Interpreting”, *The Translator* 3, no. 1: 39–54.

---- (1999) “Interpreters' Treatment of Discourse Markers in Courtroom Questions”, *Forensic Linguistics* 6, no. 1: 57–82.

---- (2002) “How Faithfully Do Court Interpreters Render the Style of Non-English Speaking Witnesses' Testimonies? A Data Based Study of Spanish-English Bilingual Proceedings”, *Discourse Studies* 4, no. 1: 25–48.

---- (2008) “Controversies over the Role of the Court Interpreter” In *Crossing Borders in Community Interpreting. Definitions and Dilemmas*, Carmen Valero-Garcés and Anne Martin (eds), Amsterdam, John Benjamins: 99–122.

Halverson, Sandra (1998) “Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study”, *Meta* 43, no. 4: 494–514/1–22.

Heritage, John (1997) “Conversation Analysis and Institutional Talk: Analyzing Data” in *Qualitative Research: Theory, Method and Practice*, David Silverman (ed), London, SAGE: 161–182.

Kadric, Mira (1999) “Interpreting in the Austrian Courtroom”, in *The Critical Link 2: Interpreters in the Community,* Roda P. Roberts, Silvana E. Carr, Diana Abraham and Aideen Dufour (eds), Amsterdam, John Benjamins: 154–164.

Lane, Chris, Katherine McKenzie-Bridle, and Lucille Curtis (1999) “The Right to Interpreting and Translation Services in New Zealand Courts”, *Forensic Linguistics* 6, no.1: 115–136.

Mikkelson, Holly (1998) "Towards a Redefinition of the Role of the Court Interpreter”, *Interpreting* 3, no. 1: 21–45.

Montalvo, Margarita (2001) “Interpreting for Non-English-speaking Jurors: Analysis of a New and Complex Responsibility”, in *ATA Proceedings for the 42nd Annual Conference*: 167–176.

Moreno Sandoval, Antonio, and José María Guirao (2006) “Morphosyntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation”, in *Spoken Language Corpus and Linguistic Informatics*, Yuji Kawaguchi, Susumu Zaima and Takagaki Toshihiro (eds), Amsterdam, John Benjamins: 199–218.

Morris, Ruth (1999) “The Gum Syndrome: Predicaments in Court Interpreting”, *Forensic Linguistics* 6, no.1: 6–29.

Nicholson, S. Nancy, and Bodil Martinsen (1997) “Court Interpretation in Denmark”, in *The Critical Link: Interpreters in the community*, Silvana Carr, Roda P. Roberts, Aideen Dufour and Ludmila Stern (eds), Amsterdam, John Benjamins: 259–270.

Niska, Helge (1995) “Just Interpreting: Role Conflicts and Discourse Types in Court Interpreting”, in *Translation and the Law*, Marshall Morris (ed), Amsterdam, John Benjamins: 293–316.

O’Connell, C. Daniel, and Kowal Sabine (1994) “Some Current Transcription Systems for Spoken Discourse: A Critical Analysis”, *Pragmatics* 4, no.1: 81–107.

Onos, Liudmila (2014) *La interpretación en el ámbito judicial: el caso del rumano en los tribunales de Barcelona*, PhD diss., Universitat Autònoma de Barcelona. URL: **http://hdl.handle.net/10803/285160** (accessed 9 June 2016).

Orozco-Jutorán, Mariana (2017a) “Efficient Equivalent Search at Your Fingertips – The Specialized Translator's Dream”, *Meta* 62, no. 1: 137–154.

---- (2017b) “Anotación textual de un corpus multilingüe de interpretación judicial a partir de grabaciones de procesos penales reales", *Revista de Llengua i Dret, Journal of Language and Law* 68: 33-56.

Ortega Herráez, Juan Miguel (2006) *Análisis de la práctica de la interpretación judicial en España. El intérprete frente a su papel profesional*. PhD diss., Universidad de Granada. URL: **http://hdl.handle.net/10481/977** (accessed 9 June 2016).

---- (2011) *Interpretar para la justicia,* Granada, Comares.

Prieto Ramos, Fernando, and Mariana Orozco-Jutorán (2015). “De la ficha terminólogica a la ficha traductólogica: hacia una lexicografía al servicio de la traducción jurídica”, *Babel* 61, no. 1: 110–130.

Rapley, Tim (2007) *Doing Conversation, Discourse and Document Analysis,* London, Sage.

Rigney, C. Azucena (1999) “Questioning in Interpreted Testimony”, *Forensic Linguistics* 6, no.1: 83–108.

Schmidt, Thomas (2011) “A TEI-based Approach to Standardising Spoken Language Transcription”, *Journal of the Text Encoding Initiative* 1: 1–22.

Schmidt, Thomas, and Kai Wörner (2009) “EXMARaLDA – Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research”, *Pragmatics* 19, no. 4: 565–582.

---- (2012). “Introduction” in *Multilingual Corpora and Multilingual Corpus Analysis*, Thomas Schmidt and Kai Wörner (eds), Amsterdam, John Benjamins: ix–xi.

---- (2014) **EXMARaLDA**. In *Handbook on Corpus Phonology*, Ulrike Gut Jacques Durand and Gjert Kristoffersen, (eds), Oxford, Oxford University Press: 402–419.

Stern, Ludmila (1995) “Non-English Speaking Witnesses in the Australian Legal Context: The War Crimes Prosecution as a Case Study”, *Law/Text/Culture* 2: 6–31.

Tusón, Amparo (1997) *Análisis de la conversación*, Barcelona, Ariel.

Vargas-Urpi, Mireia (2012) *La interpretació als serveis públics i la mediació intercultural amb el col·lectiu xinès a Catalunya.* PhD diss., Universitat Autònoma de Barcelona. URL: **http://hdl.handle.net/10803/96486** (accessed 9 June 2016).

Vargas-Urpi, Mireia, and Marta Arumí (2014) “Estrategias de interpretación en los servicios públicos en el ámbito educativo: estudio de caso en la combinación chino-catalán” *InTRAlinea*, Vol. 16. URL: **http://www.intralinea.org/current/article/estrategias_de_interpretacion_en_los_servicios_publicos_en_el_ambito_edu** (accessed 10 June 2016).

Wadensjö, Cecilia (1998) *Interpreting as Interaction*, New York, Longman.

Notes

[1] The Directive 2010/64/EU of the European Parliament and of the Council of 20 October 2010 on the right to interpretation and translation in criminal proceedings and the Directive 2012/13/EU of the European Parliament and of the Council of 22 May 2012 on the right to information in criminal proceedings

[2] Our translation, taken from the text of the law: *Ley Orgánica 5/2015, de 27 de abril por la que se modifican la Ley de Enjuiciamiento Criminal y la Ley Orgánica 6/1985, de 1 de julio, del Poder Judicial, para transponer la Directiva 2010/64/UE, de 20 de octubre de 2010, relativa al derecho a interpretación y a traducción en los procesos penales y la Directiva 2012/13/UE, de 22 de mayo de 2012, relativa al derecho a la información en los procesos penales.* [https://www.boe.es/boe/dias/2015/04/28/pdfs/BOE-A-2015-4605.pdf]

[3]The official name of the project is ‘Translation quality as a guarantee of criminal proceedings. Development of technological resources for court interpreters in Spanish-Romanian, Arabic, Chinese, French and English language pairs’ and it has been funded by the Spanish Ministry of Economy and Competitiveness (FFI2014-55029-R). Seven researchers make up the research team: Dr. Marta Arumí, Dr. Anna Gil Bardají (Universitat Autònoma de Barcelona), Dr. Anabel Borja (Universitat Jaume I), Dr. Mireia Vargas-Urpí (Universitat Pompeu Fabra) and Dr. Francisco Vigier (Universidad Pablo de Olavide) and the two main researchers who lead the team are Dr. Carmen Bestué and Dr. Mariana Orozco-Jutorán (Universitat Autònoma de Barcelona).

[4] For a thorough description of the tools, see Schmidt and Wörner (2009, 2012 and 2014).

[5] To fully understand this decision one has to bear in mind that transcribing one minute of live trial involves at least 30 minutes of work for a trained transcriber.

[6] In this respect, see, for instance, Bendazzoli (2010), Edwards (2001), Edwards and Lampert (1993), Emerson (1996), Fairclough (2001), Falbo (2005), Fowler (2007), Halverson (1998), Heritage (1997), Moreno and Guirao (2006), O’Connell and Kowal (1994), Rapley (2007), Schmidt (2011), Tusón (1997).

[7] http://www.exmaralda.org/en/

[8] To see a quick review of these and other annotation systems, visit, for instance, **http://ucrel.lancs.ac.uk/annotation.html**

[9] The results of the Law10n research project, funded by the Spanish Ministry of of Economy and Competitiveness, can be accessed at **http://lawcalisation.com/**

"The TIPp project Developing technological resources based on the exploitation of oral corpora to improve court interpreting", *inTRAlinea* Special Issue: New Findings in Corpus-based Interpreting Studies.

Stable URL: http://www.intralinea.org/specials/article/2316