EP-Poland: Building A Bilingual Parallel Corpus For Interpreting Research

Magdalena Bartłomiejczyk [a]
magdalena.bartlomiejczyk@us.edu.pl
University of Silesia in Katowice, Poland

Ewa Gumul [b]
ewa.gumul@us.edu.pl
University of Silesia in Katowice, Poland

Danijel Koržinek [c]
danijel@pjwstk.edu.pl
Polish-Japanese Academy of Information Technology, Poland

ABSTRACT

This paper reports on the process of building the EP-Poland corpus and on the first empirical applications thereof. This extensive bidirectional English-Polish corpus of original parliamentary contributions paired with professional simultaneous interpretations includes 11 European Parliament debates held between January 2016 and February 2020. The main topic of these debates is the rule of law crisis triggered by the Law and Justice government in Poland. The corpus contains over 157,000 tokens and about 20 h 45 min of recordings, counting both source and target texts. The two interpreting directions (English-Polish and Polish-English) are represented almost evenly. The annotation of the corpus completed so far includes mark-up information, POS tagging, labelling disfluency phenomena, and all forms of explicitating shifts. Manual annotation for personal deixis is in progress. An additional interesting feature is the speaker identification performed employing the X-vector method, which allowed us to identify 36 interpreters. We begin with an overview of the existing interpreting corpora. Then we proceed to explain the design features of the EP-Poland and report on two completed empirical studies analysing idiosyncratic interpreting behaviour. We conclude by outlining future development pathways and offering some remarks on corpus significance and its limitations.

Keywords: interpreting corpus; parallel corpus; simultaneous interpreting; political discourse; parliamentary interpreting

INTRODUCTION

Using corpus linguistics (CL) tools is more widespread and has a longer tradition in research on written translations than in Interpreting Studies. Baker (1993) triggered a big wave of empirical research focused, first of all, on exploring the particular properties of translated texts that distinguish them from non-translated ones (e.g., Laviosa, 1998). This may be done either by comparing translations with their originals, i.e., through parallel corpora, or by comparing translations with texts of the same type originally composed in the target language, i.e., through monolingual comparative corpora. The use of the latter, however, has led to controversies, with some scholars (e.g., Bernardini and Zanettin, 2004) arguing that the source texts on which translations are based should not be ignored. In Interpreting Studies, the interest in building

[a] *Main author*
[b] *Corresponding author*
[c] *Corresponding author*

eISSN: 2550-2131
ISSN: 1675-8021

and analysing corpora was awoken by Shlesinger (1998), and its bloom in the recent decade has been reflected, inter alia, in two comprehensive collective volumes: Straniero Sergio & Falbo (2012); Russo et al. (2018).

In this paper, we report on building the EP-Poland interpreting corpus. This is work in progress. Although the corpus is ready for many research applications, it may be still supplemented with more material and new features, as the research needs dictate. To provide the necessary background, we first briefly present CL as a research paradigm in Interpreting Studies and outline the existing interpreting corpora, in particular those that bear affinity to ours. Afterwards, we discuss in detail the subsequent stages in the development of the EP-Poland, including its design, transcription, annotation, and speaker identification. We also provide information on the first empirical studies based on specific subcorpora extracted from the EP-Poland that are currently under review. Finally, we report on the ongoing empirical endeavours as well as our research plans for the nearest future and sketch envisaged paths of further development.

CORPUS LINGUISTICS AND INTERPRETING STUDIES

Corpora are often seen simply as a resource for linguists, but their role is actually more complex. As pointed out by Saldanha & O'Brien, "CL is considered a research paradigm in its own right", as "doing research using corpora generally entails some basic assumptions as to what the object of enquiry is and how it should be studied" (2013, p. 56). In Interpreting Studies, the current widespread use of corpora reflects a major paradigm shift from experimental to observational, product-oriented research. While the former has by no means been abandoned, the latter is providing novel viewpoints and contributing to a much broader overall panorama, encompassing conference interpreting as well as, increasingly, also other interpreting modes that used to be underresearched. Reliance on a large amount of authentic data strongly embedded in its communicative context is supposed to limit the researcher's subjectivity and to guarantee ecological validity that is typically missing in experimental studies. In the words of Daniel Gile, experimental research has little explanatory power in the eyes of professional interpreters, who feel that

> important determinants of the interpreter's behaviour are only found in the 'real' professional situation, including a sense of professional responsibility, the awareness of certain expectations from colleagues and listeners, visual and other feedback from the clients and the floor as well as visual and other input from the interaction between the floor and the speakers and within the floor (2000: , p. 102)

Interpreting corpora are tiny in comparison with general language corpora and even with translation corpora. Among monolingual parliamentary corpora, for instance, the Polish Parliamentary Corpus contains 300 million tokens (Ogrodniczuk & Nitoń, 2020), while the Malaysian Hansard Corpus exceeds 157 million (Mat Awal et al., 2019). Building interpreting corpora of comparable size is unfeasible because "interpreting combines two features that have traditionally hindered the development of corpus resources: orality and interlinguistic mediation" (Bernardini et al., 2018, p. 22). Confusingly, interpreting researchers often apply the term "corpus" rather loosely, i.e., also for collections of interpretations obtained in laboratory settings (e.g., Gumul, 2017; 2021). Obviously, these do not meet the criteria set out by corpus linguists, who define a corpus as a "computerized collection of authentic texts, amenable to automatic or semi-automatic processing or analysis" (Tognini-Bonelli, 2001: 55). The minimum size from which a set of interpretations should reasonably be referred to as a "corpus" is also dubious, e.g., Liontou (2013) uses this term for three interpretations (each about 5 minutes long) paired with their source texts.

eISSN: 2550-2131
ISSN: 1675-8021

While CL is generally associated with quantitative studies relying heavily on automatic analyses, in Interpreting Studies, manual analysis of data tends to play a significant role (cf. Bendazolli 2018, pp. 5-6), and the qualitative component is relatively strong, especially in discourse-analytic research (e.g., Beaton 2007; Bartłomiejczyk 2016).

EXISTING INTERPRETING CORPORA

Plenary sessions of the European Parliament (EP) are frequently tapped for interpreting corpora primarily because of the wide range of source and target languages (24 official languages of the EU), and availability for download from the EP website in the form of MP4 videos and "verbatim reports" (transcripts of original speeches). Plenary contributions total about 430 hours per year (European Parliament 2013) and the range of topics is very wide, covering practically all the areas subject to EU legislation as well as more general, formal speeches delivered, for instance, during sessions celebrating important anniversaries, or to welcome invited guests. The idea behind such a huge number of official languages is to enable everybody to speak/listen in their native language, but in practice many participants choose a different language of which they do not have a perfect mastery, i.e., a *lingua franca* of wider diffusion (typically English).

The first large corpus fulfilling the usual CL criteria as outlined in the previous section, European Parliament Interpreting Corpus (EPIC), was created at the University of Bologna at Forlì (cf. Monti et al., 2005). It is a parallel corpus, i.e., containing original speeches and their interpretations, initially in English, Spanish and Italian. Besides in-depth methodological considerations setting many standards for future projects of this type, EPIC generated a number of diverse empirical studies. In one of the earliest, Russo et al. (2006) investigate lexical density (the ratio of content words to grammatical words) and lexical variety (the ratio of high frequency words to low frequency words), comparing interpretations into Spanish with STs originally delivered in Spanish. Unlike in previous research on written translations, lexical density turned out to be slightly higher (by about 0.5%) for interpretations. Interpretations from English were characterized by considerably higher lexical variety than interpretations from Italian. The unexpected results for lexical density made the authors also compare the interpretations with their STs, which turned out to be more lexically dense (nearly 3% for English STs and nearly 5% for Italian STs).

Bendazzoli et al. (2011) focus on disfluencies (mispronounced and unfinished words), hypothesising that under the constraints of simultaneous interpreting more disfluencies should be produced by interpreters than by original speakers and a higher proportion of them would remain uncorrected. By and large, this hypothesis is confirmed, while some exceptions point to language-pair specific effects, i.e., lower interference when interpreting between non-cognate languages. Spinolo & Garwood (2010) explore how interpreters deal with various types of metaphors. Their findings indicate that catachreses and idioms tend to be paraphrased, whereas live metaphors tend to be rendered literally. No consistent pattern emerges for metaphorical concepts (such as the EU as a building). EPIC was even used by students for their graduation theses (cf., Russo 2010), out of which the ones dealing with pragmatic aspects of interpreting (face threats and interpersonal features) seem the most interesting.

More recently, EPIC has transformed into European Parliament Translation and Interpreting Corpus (EPTIC) with the addition of new languages and another modality, i.e., translations of verbatim reports (see, e.g., Ferraresi et al., 2018). This allows for comparisons of interpretations and translations of the same STs.

Another extensive corpus was developed by a team at the University of Ghent: European Parliament Interpreting Corpus Ghent (EPICG, EPIC Ghent), including material in French, Spanish, English and Dutch. This is undoubtedly the corpus that has spawned the most

numerous empirical studies recently, of which, due to space constraints, we are able to mention but a few. Magnifico & Defrancq (2016) analyse mitigated and unmitigated face-threatening acts in English and Dutch interpretations from English with regard to interpreters' gender. While mitigation is frequently undertaken by interpreters, the results indicate that male interpreters carried out more facework to tone down unmitigated face-threatening acts produced by original speakers. Defrancq & Plevoets (2018) propose to measure cognitive load involved by means of the occurrence rate of the disfluency *uh(m)*, comparing interpretations into Dutch with original texts in the same language. They conclude that the frequency of *uh(m)* in interpretations shows a positive correlation with the ST lexical density, and a negative correlation with formulaicity in both ST and TT. Collard & Defrancq (2019), in turn, measure the ear-voice span (EVS), i.e., the time lag between the speaker and the interpreter, in six language pairs to see whether it reflects postulated cognitive differences between men and women. While the study reveals no gender differences, it identifies delivery rate, TT disfluencies and the languages involved as relevant predictors of EVS.

The total sizes of EPIC and EPICG are difficult to determine, as researchers report on various stages of their development and use different sets of subcorpora. Spinolo & Garwood (2010), for example, analyse an EPIC version in which three source languages account for nearly 8 hours of ST material. Magnifico & Defrancq (2016), in turn, report that their EPICG (with two source and two target languages) contains over 220,000 tokens.

We are also aware of an ongoing project somewhat similar to ours, Polish Interpreting Corpus (PINC) (Chmiel et al., forthcoming), a well-balanced bidirectional English-Polish parallel corpus developed primarily with the aim of investigating cognitive mechanisms of simultaneous interpreting, namely activation and inhibition. PINC contains over 170,000 tokens and includes material from plenary sessions held in 2009 and 2010: texts ranging between 100 and 500 words, ad-libbed or read out, on a wide variety of topics such as justice, agriculture, environment or health. Original speakers are only Members of the European Parliament (MEPs) using their native language, in this case, either English or Polish (i.e. contributions delivered in English as a foreign language are not included). The interpreters are all native speakers of Polish.

Importantly, also individual researchers sometimes compile large interpreting corpora with EP plenary speeches, e.g., Kajzer-Wietrzny's Translation and Interpreting Corpus (TIC) (2018), with over 250,000 tokens. TIC, unlike EPIC, EPICG or PINC, is a monolingual comparative corpus containing English interpretations and translations of speeches in French, Spanish, Dutch and German.

There also exist interpreting corpora from other settings, involving other languages than the official 24 of the EU, e.g., Chinese-English (Gu & Tipton, 2020) or Russian-English (Dayter, 2018). The former is a large corpus of annual press conferences held by Chinese Prime Ministers for English-speaking journalists, which are interpreted in the consecutive mode. Gu & Tipton (2020) use discourse analysis to reveal how Chinese interpreters boost the ideology of the "Beijing discourse" through frequent addition of self-referential items such as *we, our, China* or *government*, which is construed as "active interpreter alignment" (2020, p. 420). Dayter's Simultaneous Interpreting Russian-English corpus (SIREN), in turn, combines simultaneous interpretations from the United Nations with ones broadcast by TV channels. Dayter (2018) compares the interpretations with their STs in terms of the postulated "translation universals", concluding that interpretations into Russian conform to the predictions of translation theory, while those into English show an opposite trend.

eISSN: 2550-2131
ISSN: 1675-8021

EP-POLAND CORPUS

CORPUS DESIGN

The idea originated from the first author in late 2019. The corpus was initially intended to be a "do-it-yourself and *keep-it-for-yourself*" (Bendazzoli, 2018, p. 7) endeavour for discourse-analytic explorations basically in the vein of her earlier studies carried out on smaller and less versatile interpreting corpora (e.g., Bartłomiejczyk, 2016 analysing face-threatening acts in speeches by British Eurosceptic MEPs; Bartłomiejczyk, 2020 focusing on racist discourse by Polish MEP Janusz Korwin-Mikke), possibly with the added value of insights generated by simple corpus linguistics tools such as word-frequency lists (Corpus-Assisted Discourse Studies as outlined by Partington et al., 2013). Based on the first author's strongest languages and her interests in directionality (interpreting into one's native language vs. into a foreign language) and political discourse, it was supposed to be a bidirectional English-Polish corpus of EP plenary texts on topical political issues, containing a comparable amount of material for each interpreting direction. To ensure maximum representativeness, it should encompass a wide range of speakers, i.e., MEPs from various political Groups and Member States, representatives of the Commission and the Council, and guests. Importantly, pragmatically-oriented analyses require that discourse samples be strongly embedded in their situational context, which favoured the option of including all English and Polish contributions to a limited number of debates plus their interpretations (as, e.g., in Beaton, 2007) rather than extracting contributions meeting stricter inclusion criteria from numerous debates (as, e.g., in Chmiel et al., forthcoming). The emerging questions were, firstly, which debates to choose, and, secondly, how much material to include to strike a balance between representativeness and workload.

The first question was considered in relation to the source languages. While English (predominantly non-native) is omnipresent in practically every debate, Polish accounts for approximately 7-8% of the total speaking time (European Parliament, 2013). We considered "Poland" in the title of a debate a good predictor of content likely to attract many speakers of Polish. The legislatures 2014-2019 and 2019-2024 abounded in such debates, which results from the rule of the Law and Justice (PiS) party in Poland since 2015. PiS has raised concerns primarily due to its highly controversial reforms of the judiciary. While PiS argues that these reforms are aimed at disposing of the ubiquitous post-communist remnants, its opponents lament the loss of judicial independence and the gradual politicization of judiciary institutions such as the Constitutional Tribunal (for details on the rule of law crisis, see, e.g., Matczak, 2020). In the European Parliament, the rule of law crisis is manifest in very strong polarization of participats in the relevant debates. Nearly all of the speakers clearly fall into two opposing camps, either supporting the position that the democratically elected government of Poland has the right to do as it sees fit and the EU should not interfere (as the reforms concern areas outside the EU's common policies), or the position that the government is a danger to democracy and its actions should be thwarted by the EU (as, when joining the EU, Poland has undertaken to respect the rule of law, separation of powers, civil rights and other democratic principles).

As for the size, we decided to start with the first relevant debate in January 2016 and to try to process every debate on "Polish issues" until the end of 2020. This plan was slightly modified as a result of the Covid-19 pandemic, namely, to include only the debates before the pandemic, as its outbreak brought crucial procedural changes, e.g., MEPs participating remotely and interpreters not sharing the same booth. We did not want to account for these new variables[d].

[d] *The EP debated specifically on Poland in the hybrid mode twice by the end of 2020, dealing with "LGBTI-free zones" on 14-09-2020, and with abortion rights on 25-11-2020.*

eISSN: 2550-2131
ISSN: 1675-8021

The eleven debates finally selected are listed in Table 1. Eight are directly related to the rule of law crisis. Additionally, two were triggered by other contentious moves by PiS: attempts to tighten the existing abortion ban (no. 3), and to criminalize sexual education of children (no. 9). A debate on the future of Europe with the Polish PM also incited numerous references to the current situation in Poland.

TABLE 1. Overview of the EP-Poland corpus

	Date	**Title**	**No. of tokens**	**No. of Polish STs**	**Duration of Polish STs**	**No. of English STs**	**Duration of English STs**	**Duration of Polish and English STs**	**Approx. duration of the debate**
1	2016-01-19	Situation in Poland	23,684	14	58 min 46 sec	10	36 min 14 sec	1 h 35 min 00 sec	2 h 40 min
2	2016-09-13	Recent developments in Poland	8,654	5	19 min 08 sec	5	16 min 27 sec	35 min 35 sec	50 min
3	2016-10-05	Women's rights in Poland	12,106	28	29 min 25 sec	18	17 min 31 sec	46 min 56 sec	1 h 44 min
4	2016-12-14	Situation of the rule of law and democracy in Poland	14,337	20	27 min 20 sec	10	26 min 28 sec	53 min 48 sec	1 h 23 min
5	2017-11-15	Situation of the rule of law and democracy in Poland	15,457	13	23 min 41 sec	19	38 min 11 sec	1 h 01 min 52 sec	1 h 39 min
6	2018-02-28	Commission decision to activate Article 7 (1) TEU as regards the situation in Poland	9,257	3	09 min 34 sec	7	25 min 47 sec	35 min 21 sec	40 min
7	2018-06-13	Independence of the judiciary in Poland	12,007	5	17 min 18 sec	7	29 min 14 sec	46 min 32 sec	55 min
8	2018-07-04	Debate with the Prime Minister of Poland, Mateusz Morawiecki, on the Future of Europe	32,567	11	1h 17 min 50 sec	17	51 min 50 sec	2h 09 min 40 sec	2 h 38 min
9	2019-10-21	Criminalization of sexual education in Poland	6,342	13	15 min 36 sec	7	12 min 10 sec	27 min 46 sec	51 min

10	2020-01-15	Ongoing hearings under article 7(1) regarding Poland	12,405	16	16 min 40 sec	21	31 min 43 sec	48 min 23 sec	1 h 16 min
11	2020-02-11	Ongoing threat for the rule of law in Poland	10,318	10	13 min 06 sec	15	28 min 29 sec	41 min 35 sec	1 h 08 min
Total			157,134	138	5 h 08 min 24 sec	136	5 h 14 min 04 sec	10 h 22 min 28 sec	15 h 44 min

Consequently, the initial criteria governing the design of our corpus did not include precise preconditions such as the final number of tokens or length of individual texts. The resulting total size of the present corpus is over 157,000 tokens and about 20 h 45 min counting both source texts (STs) and target texts (TTs), which, in our view, at least sets it on par with the EPIC and EPICG projects considering the stage of development and the size of involved teams (while we have undoubtedly been able to proceed faster thanks to the fact that many important methodological questions had already been posed and considered by these pioneers of interpreting corpora). STs account for 84,632 tokens, i.e., 53.86% of the corpus, and TTs for 72,502 tokens, i.e., 46.14% of the corpus. This unequal distribution roughly reflects the overall text compression rate in the process of interpreting, which amounts to 14.33% in EP-Poland.

The achieved balance between the English-Polish and Polish-English interpreting directions has surpassed our expectations: 5 h 14 min vs. 5 h 08 min, counting only ST material. In terms of tokens, English STs account for 47,428, and Polish STs – for 37,204. This seems much less balanced than the length of the recordings, however, the difference stems from the fact that in Polish words tend to be longer and there are considerably fewer function words (e.g., no articles). The share of other source languages (and, therefore, of speeches that had to be excluded) is relatively low: on average, 65.9% of the total time of the debates accounts for STs in English (33.3%) and Polish (32.7%). The average shares calculated over a period of 3.5 years amount, respectively, to 29.1% and 7.7% (European Parliament, 2013).

The pool of original speakers includes 97 persons (55 speaking English, 38 Polish, and 4 both languages). The average length of one intervention is about 2 min 16 sec, ranging from blue card questions and answers of a few seconds to three exceptionally long speeches (34 min 55 sec, 22 min 36 sec, and 21 min 38 sec) by Polish PMs Morawiecki and Szydło. Short speaking times are imposed by strict time allocation rules, and, as a result, some speakers try to squeeze as much content as possible into their brief contributions by adopting extremely fast delivery rates. The average speaking speed is 151 words per minute (wpm) for English STs and 121 wpm for Polish STs, typical for EP plenary debates (cf. Monti et al., 2005; Bartłomiejczyk, 2016), but, at the same time, much higher than might be considered comfortable for simultaneous interpreting[e]. The modes of delivery represent the whole range of options used by EP speakers and include ad-libbing, reading out, and a mixture of the two.

[e] *The comfortable speed is about 90-120 wpm for an English ST (Pöchhacker, 2004: 129-130) and about 80-90 wpm for a Polish ST (Gumul 2017: 108).*

eISSN: 2550-2131
ISSN: 1675-8021

TRANSCRIPTION AND METADATA

Taking advantage of a research methodologies in Translation Studies seminar held during the winter semester of the academic year 2019-2020 at the University of Silesia in Katowice, a large part of transcription work was assigned to students as a task enabling them to learn first-hand about CL as one of the key topics on the curriculum. Moreover, we hoped to stimulate their reflection on simultaneous interpreting carried out by highly skilled professionals. The seminar is meant to assist third-year students in finishing their BA projects and in drafting their MA theses (if they wish to continue in the translation and interpreting programme at the graduate level). The students were invited to use the ready-made corpus or parts thereof for their MA theses. By now, two students have decided to take this opportunity, and their theses are to be completed by September 2022.

In January 2020, 57 students received fragments of verbatim reports downloaded from the EP website (about 1200 words each), which they were asked to check against delivery and complement with the corresponding interpretations, transcribed manually. STs and TTs were to be aligned in a rudimental manner by placing the equivalent paragraphs opposite of each other in the left vs. right column of a table (see Table 2). Verification involved correction to include features such as false starts, grammatical errors, contracted forms, etc. (as these are routinely deleted from EP verbatim reports – see Ferraresi et al., 2018: 724–725). We opted for orthographic naturalised transcription with the sentence boundaries marked with punctuation. The submitted transcripts were checked by the first author and their quality was highly diversified: from ones that hardly needed any correction to one student's own translation of the STs. The rest of the material, i.e., parts of earlier debates not assigned to students and the whole debates no. 10 and 11 (which were held later), was verified and transcribed by the first author and checked by the second author.

TABLE 2. Example of a speech transcription

Barbara Kudrycka (PPE), blue card answer, 0 min 16 sec, b. 08:41, 25 words	**F, 26 words**
Panie Przewodniczący! Dzisiaj Parlament <Eurobej-> Europejski nie reguluje prawnie sprawy <pol-> sprawy kobiet w Polsce. Nie robimy tego. Ale dyskutować może, o czym chce. Dziękuję.	President, <---> today, <---> the European Parliament cannot legislate in this area. We're not gonna be doing that, but we can have a debate on anything we'd like.

The metadata at this stage includes the document mark-up, such as the name of the original speaker, his/her Political Group in the EP, type of speech (i.e., a contribution on behalf of the Council, the Commision, or a particular Group, a blue card question or answer, contribution by the chairing President), duration of the ST, a starting point in the recording, number of words (ST left, TT right), and interpreter's sex.

ANNOTATION

In addition to the mark-up information described above, the corpus has been annotated for various features in line with the research interests of the authors and task-specificity of the corpus. However, we have also retained a version that follows 'clean-text policy' used, among others, in building BEC (Business English Corpus) described by Nelson (2010). Keeping the text unprocessed and clean of tags should facilitate further manual annotation of any features difficult or impossible to code automatically. The clean-text version will also be useful for any additional qualitative analysis of the material. The annotated version includes embedded linguistic tags following the XML standard, as it allows for the insertion of custom-made tag sets.

eISSN: 2550-2131
ISSN: 1675-8021

The first step of linguistic annotation, introduced at the stage of manual transcription, involved marking of disfluency phenomena (this annotation is already visible in Table 2). As essential features of orality, they form an integral part of any spoken corpora. The three types of disfluencies coded in the corpus are: hesitation markers, false starts, and anomalous pauses (roughly corresponding to the disfluencies analysed in Gumul, 2021). The first category, coded as <@>, comprises non-lexical fillers mainly in the form of prolonged vowels, i.e., the so-called filled pauses. As false starts we classified retraced and non-retraced truncations at the word level, coded with a hyphen following an interrupted word, e.g., <pol->. Anomalous pauses (coded as <--->) comprise only pauses exceeding 3 seconds. Such a high threshold should unambiguously point to non-strategic interruptions of a speech flow, possibly indicating processing problems.

The second layer of linguistic annotation was POS tagging conducted with the aid of the automatic taggers (spaCy toolkit for the texts in English and Concraft-pl for the Polish subcorpus). Selected samples of both subcorpora were manually post-processed by two researchers to verify the accuracy of annotation. The accuracy rate was better for tagging in English (over 98%). POS tagging of morphologically rich languages such as Polish is a more challenging task due to inflections. Therefore, we opted for the Concraft-pl tagger, which scores much higher in accuracy than other currently available taggers for Polish such as Pantera, OpenNLP or WMBT (see, e.g., Krasnowska-Kieraś & Kobyliński, 2019). We achieved a reasonable accuracy of around 92%, which required us to perform manual post-processing of the data in Polish. The inter-coder agreement in manual post-processing was 100%.

The next stage of annotation involves manual tagging for certain features not extractable by means of software. In line with our current research interests, we have first completed adding tags for all types of explicitating shifts. The annotation of EP-Poland for this feature departs from those adopted in previous corpus studies of explicitation (see, e.g., Kajzer-Wietrzny, 2012). Since we focus on explicitation as a shift between source and target text rather than only target text explicitness, the annotation requires comparative analysis of STs and TTs, while tags are only placed in the TTs. Purely manual annotation is also necessary due to the scope of analysed shifts. Our intention was to investigate the entire spectrum of surface forms of explicitation ranging from cohesion-related surface additions or specifications (adding connectives – coded as ACon, intensifying cohesive ties – ICT, lexicalising pro-forms – LxPF, filling out elliptical constructions – FEll, reiterations of lexical items – Reit) through syntactic transformations (replacing nominalisations with verb phrases – NVP) to other texture-enriching shifts (adding modifiers and qualifiers – M/Q, inserting hedges and discourse organising items – Hdg, DOI, including explanatory remarks – ExR, providing full expression for acronyms or abbreviations – FAA disambiguating lexical metaphors – DLM, performing shifts involving proper names – PrN, lexical specification – LxSp, and meaning specification – MSp). This annotation is illustrated in Table 3. The first study making use of it (Gumul & Bartłomiejczyk, under review) will be described in some detail later on.

TABLE 3. Example of annotation for expliciation

Ale przyjeżdżam również tutaj, przyjeżdżam tutaj również, proszę państwa, bo mam głębokie poczucie odpowiedzialności nie tylko za to, co dzieje się w Polsce, ale mam głębokie poczucie odpowiedzialności za to, co dzieje się w Europie. Mówicie państwo o migrancji, o migrantach – to jest poważny problem – i państwo o tym doskonale wiecie	--- But I have come in here to this Parliament for another reason. The reason <Reit> is my deep sense of responsibility, not also not only for what is going on in Poland, but also for what is going on in Europe. So <ACon> you see, I'm talking about migration. Migration is a huge problem and you know this very well.
(…)	(…)

Szanujemy państwo prawa i szanujemy decyzje podjęte przez rząd poprzedni. Uczestniczymy w dyskusji i w procesie, który w tej chwili wypracowuje Unia Europejska w sprawie migracji. Potrafimy sobie z tym poradzić? Musimy sobie wszyscy na to państwo odpowiedzieć, i my i wy, a szczególnie tutaj, w tym miejscu, w Parlamencie Europejskim, bo tego oczekują dzisiaj od was obywatele Europy zatroskani o swoje bezpieczeństwo. Pracujmy nad tym – to jest bardzo ważne, to jest wielkie wyzwanie dla Europy. Ja nie chcę, żeby w moim kraju, żeby w Polsce, ludzie bali się, żeby narastały obawy antyeuropejskie. Zróbmy wszystko, zróbmy wszystko, by Europa rozwi-rozwijała się w spokoju i by była wspólnotą suwerennych, równych, sprawiedliwie rządzonych państw.	We are a country based on the rule of law and we will respect decisions taken by the previous government and <ACon> we will participate in the currently discussed mechanisms of migration in the EU. Can we cope with the problem of migration? <LxPF> Well, <Acon> we, all of us, we have to provide and answer to this question, especially here, in the European Parliament, because this is what @ the citizens of Europe are expecting in our from us, from you. Let's work towards this goal. <LxPF> This is a major objective for us, this is a challenge ahead of us. I don't want Poles to become @ Eurosceptic, I don't want Poles to be critical of Europe. Therefore <ACon> we should spare no effort to make Europe a community of sovereign, well-governed countries.

For the fragment provided in Table 3, the annotation shows that the interpreter has performed seven explicitating shifts. He reiterates the word "reason", which only has one occurrence in the ST. He also adds three conjunctions to explicitate the underlying causative ("so" and "therefore") and additive ("and") relations. Moreover, in the TT there is an additional continuative "now" marking an announcement of evaluative statements acknowledging a different point of view as well as two instances of lexicalisation of proforms (two target-text pronouns are translated as "migration" and "goal").

Another type of manual annotation currently under way is of pragmatic nature. It focuses on personal deixis, a prominent topic for analysts of political discourse. However, in interpreted political discourse, the use of personal deixis ultimately depends on the interpreter. The professional norm requires that the interpreter should retain the deictic perspective of the original speaker, but departures from that norm are not uncommon (as shown, e.g., in Bartłomiejczyk, 2016). At present, the first author is working on two research projects: one focused on self-reference (operationalized as first-person singular pronouns and verb forms) and the other one exploring address (operationalized as second person singular and plural pronouns and verb forms as well as nominal forms of address, e.g., honorifics). Further plans include exploring the WE-perspective, which is very widely used in political discourse and has enormous ideological significance. Unlike explicitation, these forms are present both in STs and in TTs, although specific occurrences do not necessarily correspond to each other. In other words, personal deixis may be either omitted or added by the interpreter. In the former case, it will be present only in the ST, and in the latter only in the TT. Furthermore, while corresponding personal forms may be present in the ST and the TT, the deictic perspective is sometimes shifted by the interpreter, for example, from second to third person (which may, e.g., mitigate the illocutionary force of an accusation). A comparative analysis shows that such forms correspond to each other, but they are not equivalent. All of these nuances and more need to be annotated throughout the corpus – see Table 4.

TABLE 4. Example of annotation for personal deixis

Ale jestem <FP1-Eq> tu, bo chcę <FP1/Eq> podjąć ten dialog, o którym była mowa. Chcę <FP1/Eq> opowiedzieć państwu <Add2=V/Eq> o Polsce, chcę <FP1/Eq> wyjaśnić wszystkie wątpliwości i wierzę <FP1/Eq> w to głęboko, że z dobrą wolą, z jaką się tutaj spotkam <FP1/SHIFT> będziemy mogli <FP2/Eq> po tej debacie wyjść wszyscy w przekonaniu, że oto Polska jest silnym, dobrze	But I am <FP1/Eq> here, because I want <FP1/Eq> to engage in this dialogue. I'd like <FP1/Eq> to talk to you <Add=/Eq> about Poland. I'd like <FP1/Eq> to dispel any doubts you may have <Add=/TT>. And I think <FP1/Eq> that with the good will that there is between us <FP2/SHIFT>, we can <FP2/Eq> all leave the chamber after this debate, believing that Poland is a strong country, a Member State of the

rozwijającym się członkiem Unii Europejskiej i że wszyscy jesteśmy <FP2/Eq> z tego dumni.	European Union that is developing on a positive trajectory, and we can be <FP2/Eq> proud of that.

<FP1> and <FP2> stand for first person singular and plural, respectively. <Add> stands for addressative forms, which are further classified as face-enhancing <Add+>, face-threatening <Add-> or neutral <Add=>. In contrast to the highly versatile English "you", Polish syntax makes a distinction between singular and plural forms of address; consequently, singular ones are coded as <Add1>, while plural ones are coded as <Add2>. Moreover, Polish is a T/V language, offering a choice between informal/ more intimate and formal/ more distanced forms. These are coded as <T> and <V>, respectively. The part after the slash refers to the equivalence between the ST and the TT or lack thereof. Forms that have their close counterparts are coded as </Eq>, and those which do not – as </ST> (present only in the ST) or </TT> (present only in the TT). Finally, </SHIFT> means that the deictic perspective has been modified by the interpreter. For the fragment provided in Table 4, the annotation shows that the interpreter is transferring the original deixis relatively closely. There is only one shift from first person singular to plural and one neutral form of address added by the interpreter.

IDENTIFICATION OF INDIVIDUAL INTERPRETERS

The process of automatically identifying people by the timbre of their voice is very well studied and has a long history. There are many applications of this technology in both analyzing recorded speech (e.g., datamining, speech biometrics) as well as reinforcing other technologies and tasks (e.g., speech recognition, spoken language understanding). For our corpus, the task was limited to identifying the interpreters, as the original speakers are known. Therefore, we had to divide our collection of audio files into segments each containing the voice of one person (an interpreter given a fictitious name). Usually, this problem is solved by first defining a separate set of recordings containing the voices of all the speakers, known as the enrollment set, but in our case this set had to be collected from the samples within the analysed dataset itself. An iterative process was thus formulated:

1. The dataset was initially analysed and annotated using the information from another, similar dataset (Polish interpreters' voice samples from the PINC corpus, see Chmiel et al., forthcoming);
2. Each segment was quickly evaluated manually and corrections were made: new voices with timestamps for each voice were identified "by ear";
3. The dataset was analysed from the beginning using updated information;
4. The process was repeated from step 2 until no more corrections were required.

We performed nine iterations to annotate nearly the whole dataset (approx. 0.32% of the material, mostly very short contributions, proved impossible to unequivocally ascribe to specific interpreters).

The technology used is based on the X-vector method (Snyder et al., 2018). It works by converting a short audio segment into a 128-element vector that serves as a descriptor of the segment within the speaker vector space. The standard approach, taken also here, is to use PLDA (Probabilistic Linear Discriminant Analysis) classification on every segment-speaker pair and choose the result that yields the highest score. A low score indicates that the particular segment does not match any speaker in the database.

For our project, we initially used simple energy-based voice activity detection to divide the audio into short segments that contain only speech. We then analysed each segment individually. We only needed to extract the X-vector embedding once. All the updates of the speakers were then performed using only the precomputed vectors. Each new speaker was

defined by a single timestamp denoting the beginning of their speech. A set of X-vectors computed from the span of roughly 60 seconds beginning at the timestamp was collected and averaged to generate a single X-vector defining the speaker.

The final analysis was presented in the form of a spreadsheet where each row represented an analysed speech segment and the columns contained information on identified speakers and their scores. Within the whole corpus, we identified 36 interpreters, out of whom 10 are the same as in the PINC corpus. 15 interpreters work only into Polish, 11 only into English, and 10 in both the directions. Some of those 36 interpreters, however, render only one or two speeches, so their output is hardly representative. As the EU interpreting services are generally unwilling to cooperate with researchers, we do not possess any information on the interpreters beyond the fact that they all must have passed the demanding accreditation procedure to qualify as EU interpreters (see, e.g., Graves et al., 2022).

FIRST EMPIRICAL STUDIES BASED ON THE CORPUS

As for now, two papers that make use of the speaker identification feature as decribed above have been submitted to reputable Translation Studies journals.

Gumul & Bartłomiejczyk (under review) investigate the individual differences in explicitating styles. The relevant annotation of explicitation has already been outlined here. It has been indicated in literature (e.g., Duflou, 2016) that Language Units in the EP function as close-knit communities of practice, whose members cooperate regularly with a relatively small set of colleagues and tend to adopt common interpreting strategies and default translations of certain phrases. In the light of this, we hypothesized that their explicitating styles might be convergent, i.e., exhibit limited variety. The analysis accounts for frequency (lean, abundant or extreme explicitating styles) and consistency (consistent or sporadic explicitating styles).

In order to eliminate the variables of source-text features and speaking styles of original speakers, analysing interpreting style requires a representative and varied sample of outputs by each interpreter. The entry thresholds also depend, to a large extent, on the features that researchers endeavour to investigate, as rare phenomena will only be discernible in large samples. Consequently, we needed to establish a specific threshold for inclusion into this study. Considering that plenary speeches are predominantly very short (2 min 16 sec on average) and explicitation is relativelely common, we settled on 15 minutes (counting TTs only) and at least four different speakers. Fifteen interpreters exceed this threshold, but three of them are clearly members of the English Language Unit who needed to be excluded as members of a different community of practice (they interpret from Polish as a foreign language). The outputs of the remaining 12 interpreters jointly account for over 5 h 08 min, i.e., about 50% of the whole corpus. The shortest sample is 18 min 13 sec, and the longest – 50 min 05 sec.

The results do not confirm our initial assumption, as our interpreters' explicating styles turned out to be very divergent. As far as the frequency of explicitations is concerned, the values range from 0.63 to as many as 3.64 shifts per 100 ST words, with six, four and two interpreters falling into the lean, abundant and extreme category, respectively. A clear pattern of relatively evenly distributed shifts and/or a marked preference for certain forms of explicitation was revealed for seven interpreters, whose style is therefore characterised as consistent. The remaining five interpreters' explicitating style is sporadic (typically coinciding with lean).

Bartłomiejczyk & Rojczyk (under review) explore non-native accent as an inherent feature of interpreting from the native into a foreign language. Traditionally, this interpreting direction has been dispreferred, often seen as a necessary evil rather than a quality service in its own right. Nevertheless, many interpreters working for the EP interpret into English as a foreign language as a matter of course, in particular from languages of relatively low diffusion,

such as Polish. As already indicated here, 10 interpreters in the EP-Poland Corpus work in both the interpreting directions, and all of them are native speakers of Polish. The entry threshold had to account for English TTs only. Considering that pronunciation may be assessed on the basis of relatively small samples, the threshold was established at five minutes and at least two different speeches. Eight interpreters met this criterion.

The study uses the methodology from L2 speech research, which relies on identifying segmental and suprasegmental departures from native pronunciation norms. Further annotation of the emergent subcorpus for the needs of this study was not necessary, as the recordings were inspected aurally and visually by means of spectrogram and waveform. The detected problems, unsurprisingly, resulted mostly from the differences between Polish and English sound systems. Although some of the interpreters sounded more native-like than others, all the analysed interpretations were evaluated as phonetically proficient and fully intelligible. What appears to impact the general impression and comprehensibility more negatively than the interpreters' slightly non-native pronunciation are disfluencies and at times too rapid articulation rates; problems not directly linked with non-native language production. Most interpreters' performance was consistent across the samples, however, pronunciation locally deteriorated in three. Two of them seemed to resort to articulatory habits from their native language for emotionally loaded speech. The third interpreter was affected by similar interference when his delivery rate slowed down considerably and became interspersed with numerous filled pauses, which suggests correspondence with increased production load.

FUTURE DEVELOPMENT PATHWAYS

As the rule of law crisis in Poland persists, the corpus could likely be extended in the future with new debates. In particular, we are considering including some more debates on Poland held during the covid pandemic as a separate subcorpus enabling us to compare the "regular" debates with the "hybrid" ones in terms of interpreters' strategic processing or even possible quality deterioration for speeches that are delivered remotely.

We are planning to additionally classify the texts into some sub-corpora, beyond the basic divisions according to language (English/Polish) and status as STs or TTs. For example, the fact that EP videos focus on speakers would enable us to easily distinguish between STs that are delivered impromptu, read out from script, and partly read out, partly delivered impromptu. However, for STs based on scripts, we have no possibility of knowing whether the interpreters rendering them were given the scripts beforehand, which certainly makes a huge difference for the manner of processing such a speech. We could also create two sub-corpora of texts clearly reflecting the ideological divide between the supporters of the Law and Justice government (and its right to introduce reforms in Poland on the basis of its democratic mandate) and the opponents, who demand EU intervention due to blatant violation of democratic principles. This would allow us to explore ideological shifts in the interpretations to verify whether EP interpreters reveal some ideological leanings towards one of the camps (as shown, for example, by Gu and Tipton, 2020, for Chinese interpreters).

Finally, we are also contemplating the alignment of texts with audio tracks, both monolingual (for interpretations) and bilingual (multi-layer, including both STs and TTs). The former would, for example, facilitate annotation of some prosodic features, while the latter would enable us to measure the ear-voice span for certain ST items (e.g., non-core lexemes, "false friends", etc.) or at particular predefined points, such as sentence-initial or sentence-final words. Both prosody and EVS are acknowledged as important indicators of interpreters' cognitive processing (see, e.g., Defrancq & Plevoets, 2018; Collard & Defrancq, 2019).

eISSN: 2550-2131
ISSN: 1675-8021

CONCLUSION

CORPUS SIGNIFICANCE AND LIMITATIONS

Our corpus provides fertile ground for the investigation of various phenomena related to simultaneous interpreting, far beyond the initial, modest plans. In the first place, we are focusing on explicitation, L2 pronunciation and personal deixis, as outlined in the previous sections. These empirical studies are quite divergent in terms of their topics and methodologies, which already now bears witness to the wide range of applicability of the data at our disposal.

Corpora such as the EP-Poland are particularly suited for quantitative analyses, which will permit us to identify and observe any discernible trends in interpreting behaviour of experienced professionals not identifiable in smaller corpora intended mainly for manual analysis. Another feature that we consider crucial is the extra-linguistic significance of the corpus. Coding the data for disfluencies will allow us to investigate them as indicators of cognitive processing following the line of research initiated in Gumul, 2021. The speaker identification is another advantage that only few interpreting corpora possess (e.g., TIC, PINC). It facilitates the comparison of interpreters' idiosyncratic behaviour, which, obviously, is not limited to explicitation only.

Finally, we would like to mention some limitations. EP-Poland cannot be transformed into an intermodal corpus (like EPTIC), as translation of plenary speeches was discontinued in 2011. The range of topics is rather limited, precluding specialist discourses related to realms other than justice. Other languages (e.g., German or Spanish) could be added to upgrade the EP-Poland to a multilingual corpus, but they would be represented predominantly as target languages. Nevertheless, we do hope that the ample naturalistic data constituting the corpus, and its additional features, in particular the identification of individual interpreters and annotation of phenomena beyond the usual tagging of POS, will provide various research opportunities for the present team, and, possibly, also for other interested scholars. For the time being, there are no plans to make the corpus available on-line, but colleagues who envisage joining our endeavours in the future are welcome to contact the first author.

ACKNOWLEDGEMENT

We thank the PINC team and their leader, Agnieszka Chmiel, for permission to use their voice samples in the process of interpreter identification.

REFERENCES

Baker, M. (1993). Corpus linguistics and translation studies: implications and applications. In M. Baker, Francis, G. & Tognini-Bonelli E. (Eds.) *Text and technology: in honour of John Sinclair* (pp. 233–250). Amsterdam/Philadelphia: Benjamins.

Bartłomiejczyk, M. (2016). *Face Threats in Interpreting. A Pragmatic Study of Plenary Discourse in the European Parliament.* Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Bartłomiejczyk, M. (2020). How much noise can you make through an interpreter? A case study on racist discourse in the European Parliament. *Interpreting*. 22(2), 238–261. https://doi.org/10.1075/intp.00042.bar

Bartłomiejczyk, M. & Rojczyk, A. (under review). How native-like do conference interpreters sound in L2? A phonetic analysis of retour interpretations into English in the European Parliament.

Beaton, M. (2007). *Intertextuality and ideology in interpreter-mediated communication: The case of the European Parliament.* Unpublished Ph.D thesis, Heriot-Watt University.

Bendazzoli, C. (2018). Corpus-based Interpreting Studies: Past, present and future developments of a (wired) cottage industry. In M. Russo, C. Bendazzoli & B. Defrancq (Eds.) *Making way in corpus-based interpreting studies* (pp. 1–20). Singapore: Springer.

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

Bernardini, S. & Zanettin, F. (2004). When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In A. Mauranen & P. Kujamaki (Eds.) *Translation universals: do they exist?* (pp. 51–62). Amsterdam: Benjamins.

Bernardini, S., Ferraresi, A., Russo, M., Collard, C. & Defrancq, B. (2018). Building interpreting and intermodal corpora: A *how-to* for a formidable task. In M. Russo, C. Bendazzoli & B. Defrancq (Eds.) *Making way in corpus-based interpreting studies* (pp. 21–42). Singapore: Springer.

Chmiel, A., Kajzer-Wietrzny, M., Koržinek, D., Janikowski, P., Jakubowski, D. & Polakowska, D. (forthcoming). Fluency parameters in the Polish Interpreting Corpus (PINC). In M. Kajzer-Wietrzny, S. Bernardini, A. Ferraresi & I. Ivaska (Eds.) *Empirical Investigations into the Forms of Mediated Discourse at the European Parliament*. Berlin: Language Science Press.

Collard, C. & Defrancq, B. (2019). Predictors of ear-voice span, a corpus-based study with special reference to sex. *Perspectives.* 27(3), 431–454. https://doi.org/10.1080/0907676X.2018.1553199

Dayter, D. (2018). Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *Forum.* 16(2), 241–264. https://doi.org/10.1075/forum.17004.day

Defrancq, B. & Plevoets, K. (2018). The cognitive load of interpreters in the European Parliament: A corpus-based study of predictors for the disfluency *uh(m)*. *Interpreting.* 20(1), 1–32. https://doi.org/10.1075/intp.00001.ple

Duflou, V. (2016). *Be(com)ing a conference interpreter: An ethnography of EU interpreters as a professional community.* Amsterdam/Philadelphia: Benjamins.

European Parliament. (2013). *Towards More Efficient and Cost Effective Interpretation in the European Parliament.* Retrieved January 10, 2021 from https://www.europarl.europa.eu/doceo/document/A-7-2013-0233_EN.html

Ferraresi, A., Bernardini, S., Petrović, M. M. & Lefer, M.-A. (2018). Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta.* 63(3), 717–738. https://doi.org/10.7202/1060170ar

Gile, D. (2000). Issues in interdisciplinary research into conference interpreting. In B. Englund Dimitrova & K. Hyltenstam (Eds.) *Language processing and simultaneous interpreting. Interdisciplinary perspectives* (pp. 89–106). Amsterdam and Philadelphia: Benjamins.

Graves, A., Pascual Olaguíbel, M., & Pearson, C. (2022). Conference interpreting in the European Union institutions. In M. Albl-Mikasa, & E. Tiselius (Eds.) *The Routledge handbook of conference interpreting* (pp. 104–114). London and New York: Routledge.

Gu, C. & Tipton, R. (2020). (Re-)voicing Beijing's discourse through selfreferentiality: a corpus-based CDA analysis of government interpreters' discursive mediation at

China's political press conferences (1998–2017). *Perspectives.* 28(3), 406–423. https://doi.org/10.1080/0907676X.2020.1717558

Gumul, E. (2017). *Explicitation in simultaneous interpreting: A study into explicitating behaviour of trainee interpreters.* Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Gumul, E. (2021). Explicitation and cognitive load in simultaneous interpreting: Product- and process-oriented analysis of trainee interpreters' outputs. *Interpreting* 23 (1), 45-75. https://doi.org/10.1075/intp.00051.gum

Gumul, E. & Bartłomiejczyk, M. (under review). Interpreters' explicitating styles: A corpus study on material from the European Parliament.

Kajzer-Wietrzny, M. (2012). *Interpreting Universals and Interpreting Style.* Unpublished Ph.D thesis, PhD dissertation, Adam Mickiewicz University in Poznań.

Kajzer-Wietrzny, M. (2018). Interpretese *vs.* non-native language use. The case of optional *that*. In M. Russo, C. Bendazzoli & B. Defrancq (Eds.) *Making way in corpus-based interpreting studies* (pp. 97–113). Singapore: Springer.

Krasnowska-Kieraś, K. & Kobyliński, Ł. (2019). Part of speech tagging for Polish. *Poznan Studies in Contemporary Linguistics*, 55(2), 211-237. https://doi.org/10.1515/psicl-2019-0009

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta.* 43(4), 557–570. https://doi.org/10.7202/003425ar

Liontou, K. (2013). Strategies in German-to-Greek simultaneous interpreting: A corpus-based approach. *Gramma.* 19, 37–56.

Magnifico, C. & Defrancq, B. (2016). Impoliteness in interpreting: A question of gender? *Translation & Interpreting* 8(2), 26–45. https://doi.org/10.12807/ti.108202.2016.a03

Mat Awal, N., Jaludin, A., Rahman, A. N. C. A. & Abdullah, I. H. (2019). "Is Selangor in deep water?" A corpus-driven account of air/water in the Malaysian Hansard Corpus (MHC). *GEMA Online Journal of Language Studies.* 19(2), 99–120. http://doi.org/10.17576/gema-2019-1902-07

Matczak, M. (2020). The clash of powers in Poland's rule of law crisis: Tools of attack and self-defense. *Hague Journal on the Rule of Law.* 12, 421–450. https://doi.org/10.1007/s40803-020-00144-0

Monti, C., Bendazzoli, C., Sandrelli, A. & Russo, M. (2005). Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta.* 50(4), 1079–1147. https://doi.org/10.7202/019850ar

Nelson, M. (2010). Building a written corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.) *The Routledge handbook of corpus linguistics* (pp. 53–65). London: Routledge.

Ogrodniczuk, M. & Nitoń, B. (2020). New developments in the Polish Parliamentary Corpus. In D. Fišer, M. Eskevich & F. de Jong (Eds.) *Proceedings of the Second ParlaCLARIN Workshop*, (pp. 1–4). Marseille: European Language Resources Association (ELRA).

Partington, A., Duguid, A. & Taylor, C. (2013). *Patterns and meanings in discourse. Theory and practice in corpus-assisted discourse studies (CADs)*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.55

Pöchhacker, F. (2004). *Introducing Interpreting Studies.* London: Routledge. https://doi.org/10.4324/9781315649573

Russo, M. (2010). Reflecting on interpreting practice: Graduation theses based on the EPIC. In L. Zybatow (Ed.) *Translationswissenschaft – Stand und Perspektiven* (pp. 35–50). Frankfurt am Main: Peter Lang.

Russo, M., Bendazzoli, C. & Defrancq, B. (Eds.). (2018). *Making way in corpus-based interpreting studies*. Singapore: Springer. https://doi.org/10.1080/0907676X.2019.1594127

Russo, M., Bendazzoli, C. & Sandrelli, A. (2006). Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: Extended analysis of EPIC (European Parliament Interpreting Corpus). *Forum*. 4(1), 221–254. https://doi.org/10.1075/forum.4.1.10rus

Saldanha, G. & O'Brien, S. (2013). *Research methodologies in Translation Studies.* Manchester: St Jerome.

Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta*. 43(4), 1–8. https://doi.org/10.7202/004136ar

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 5329–5333). IEEE.

Spinolo, N. & Garwood, J. (2010). To kill or not to kill: Metaphors in simultaneous interpreting. *Forum.* 8(1), 181–211. https://doi.org/10.1075/forum.8.1.08spi

Straniero Sergio, F.& Falbo, C. (Eds.). (2012). *Breaking Ground in Corpus-based Interpreting Studies.* Frankfurt am Main: Peter Lang. https://doi.org/10.1016/j.system.2014.02.010

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

ABOUT THE AUTHORS

Magdalena Bartłomiejczyk holds a PhD (2004) and a post-doctoral degree (2017) in Linguistics. She is a professor at the Institute of Linguistics of the University of Silesia in Katowice, where she currently teaches conference interpreting. Her scholarly interests include Interpreting Studies, Pragmatics, Sociolinguistics, Discourse Analysis and the newest developments in the Polish language.

Ewa Gumul holds a PhD (2004) and a post-doctoral degree (2018) in Linguistics. She is a professor at the Institute of Linguistics of the University of Silesia in Katowice, where she currently teaches MA seminars in Translation & Interpreting Studies, translation, and conference interpreting. Her scholarly interests include explicitation, interpreting style, and the method of retrospection.

Danijel Koržinek holds a PhD (2016) in Computer Science. His scholarly interests focus on speech recognition and other technologies for automating the analysis of speech data. He has been coordinating the development of speech analysis tools in the Polish branch of the Clarin project, whose goal is to facilitate research in various fields of humanities and social sciences.

eISSN: 2550-2131
ISSN: 1675-8021