

Chapter 3

Fluency parameters in the Polish Interpreting Corpus (PINC)

Agnieszka Chmiel^a, Danijel Koržinek^b, Marta Kajzer-Wietrzny^a, Przemysław Janikowski^c, Dariusz Jakubowski^c & Dominika Polakowska^a


^aAdam Mickiewicz University ^bPolish-Japanese Institute of Information Technology ^cUniversity of Silesia

The following chapter introduces PINC — the Polish Interpreting Corpus, a Polish-English and English-Polish corpus of short European Parliament speeches and their interpretations. The uniqueness of PINC, apart from its language combination, consists in careful balancing of mode of delivery, in rich metadata, interpreter identification and availability of a strictly controlled subcorpus of retour interpretations. The chapter also briefly presents custom-built tools used in the making of the corpus, especially for transcription, text-audio alignment at word level and interpreter identification. To showcase PINC's potential for analysing various aspects of simultaneous interpreting, we examined fluency parameters, such as speaking rate and pauses, in the Polish-English subcorpus. We found that interpreting speed was modulated by the source text speaking and articulation rate and the target text compression rate. Target texts had fewer but longer silent pauses and more numerous and longer filled pauses. Together with shorter runs, understood as utterances uninterrupted by pauses, this suggests more fragmented delivery of interpretations. We also found interesting individual differences in compression rate with the majority of interpreters producing interpretations longer than the source texts.

1 Introduction

New empirical paradigms require constant development of tools that would allow us to investigate increasingly challenging research questions. This is particularly visible in the case of Corpus Interpreting Studies, where new incarnations



Agnieszka Chmiel, Danijel Koržinek, Marta Kajzer-Wietrzny, Przemysław Janikowski, Dariusz Jakubowski & Dominika Polakowska. 2022. Fluency parameters in the Polish Interpreting Corpus (PINC). in Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Mediated discourse at the European Parliament: Empirical investigations*, 63–91. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977042 

of interpreting or intermodal corpora based on the European Parliament plenary debates have emerged every few years since 2005, when EPIC: European Parliament Interpreting Corpus (Monti et al. 2005) was announced. Despite the readiness of the corpus creators to collaborate and share their data (most corpora are available either online or from their owners upon request), all of them stand by their own preferred corpus tools and compilation procedures as these fit their research needs best. This is related to the fact that interpreting is not limited to the text and the linguistic aspects captured in transcripts do not reflect the full communication event. As the Corpus Interpreting Studies pioneer, Shlesinger (1998: 1), put it “[w]hile transcription, however laborious, can provide us with a representation of the interpreter’s linguistic output, its failure to reflect the concomitant paralinguistic dimensions is a major drawback”. Hence, so far, most interpreting corpora have been compiled with particular research objectives in mind. Such is also the case of PINC: The Polish Interpreting Corpus, which, at a later stage of the project, will be used to analyse activation and inhibition and thus needs intense annotation of such features as e.g. temporal details of individual words, pause length or word-level alignment. Many of these features will enable a peek into the process of interpreting, rather than being strictly product-oriented, and the data obtained will inform the selection of stimuli for final-stage experimental procedures. This puts quite a heavy demand on strict balancing and control as well as the sheer size of the corpus.

This chapter presents this newly created Polish Interpreting Corpus and offers an example study that shows the potential of PINC in analysing various aspects of simultaneous interpreting. We have decided to concentrate on interpreter fluency, including speaking rate and pauses, and to look for characteristics in the source text that modulate fluency parameters in interpretations.

To the best of our knowledge, only three interpreting corpora have been created for the Polish-English language combination so far. Two of them (Dumara 2015, Bartłomiejczyk 2016) were analysed manually and with a narrow research focus, such as intrusive pronouns or face threats. The third is a Polish-English small-scale subcorpus currently available as part of EPTIC (Department of Interpreting and Translation - Forlì Campus). We hope that PINC, thanks to its size and advanced analytical metadata (to be described below), will make it possible to tackle varied and numerous research questions.

2 PINC: a new member of the EPIC suite of corpora

2.1 Features

The Polish Interpreting Corpus (PINC) adds to an ever-growing family of interpreting or intermodal corpora derived from the European Parliament debates called by Bernardini et al. (2018) “the EPIC suite of corpora”. As summarised by Bernardini et al. (2018: 22), “[t]he availability of interpretations and translations from and into a large number of languages, the ease of access to the videos (downloadable from the Internet), and the high professional standards of the interpreters” makes this source very promising for interpreting corpora, which is why the EPIC suite is constantly growing. Next to EPIC: European Parliament Interpreting Corpus (Monti et al. 2005), TIC: Translation and Interpreting Corpus (Kajzer-Wietrzny 2012: 57), EPICG: European Parliament Interpreting Corpus – Ghent (Defrancq et al. 2015) and EPTIC: European Parliament Translation and Interpreting Corpus (Ferraresi & Bernardini 2019), PINC comprises a collection of recordings and transcripts of speeches delivered during the plenary sessions of the European Parliament, as well as their simultaneous interpretations. These were obtained from the Europarl website (Directorate-General for Communication).

Several aspects of the PINC compilation process have been modelled on the work of the creators of the other corpora of the EPIC suite (e.g. EPIC or EPTIC). Thus, the texts compiled in the PINC corpus follow the same topic classification as EPIC and EPTIC for ease of comparison; similar contextual metadata have been collected and transcription guidelines were, to a large extent, very much alike. Similarly to TIC, interpreters’ voices have been distinguished from one another and individual codes have been assigned to each voice. The uniqueness of PINC consists in the specific language combination, careful balancing of mode of delivery, detailed interpreter voice identification, speech-to-text and sentence alignment of the whole corpus and in the specific tools employed to automatise parts of the compilation process. Some of these tools and features are described in more detail below.

2.1.1 Corpus size, speech length and mode of delivery

Since we are interested in a fully bidirectional analysis, PINC consists of four balanced subcorpora: Polish source texts (ST-PL), their interpretations into English (TT-EN), English source texts (ST-EN) and their interpretations into Polish (TT-PL). All of these were collected from the Europarl website from plenary session recordings of the European Parliament sittings taking place between Jan-

uary 2009 and September 2010. The reasons for selecting such a time frame were twofold. First, we wanted to use verbatim reports to facilitate the transcription process and later recordings are not accompanied by them. Second, we are planning to correlate corpus data for individual interpreters with other data, such as working memory spans, obtained from the same interpreters in the same time frame and analysed in other studies (Chmiel 2012; Chmiel 2016; Chmiel 2018).

The PINC corpus comprises of texts ranging between 100 and 500 words, with the mean text length of 204 and the median text length of 183 words. Including a higher number of shorter speeches rather than fewer longer ones made it possible to achieve greater variation within the data, as a greater proportion of longer speeches in a corpus of the same size could have easily skewed the data. Thus, texts longer than 500 words have been excluded from the corpus altogether.

As in other corpora of the EPIC suite, speeches in PINC are annotated for mode of delivery. Following EPIC (Monti et al. 2005), most EP-based interpreting corpora use a three-way classification of mode of delivery: *impromptu* (for unscripted speeches), *read* (for scripted speeches) and *mixed* (semi-scripted speeches). In the course of compilation of PINC, a decision was made to include only the first two types of speeches in the corpus; hence speeches of varying degree of scriptedness are not part of the PINC corpus. The reason for excluding mixed speeches was that we found it difficult to indicate precise and objective criteria for assigning speeches to that category. Table 1 presents basic data about the number of speeches and tokens, as well as speech rate in each subcorpus of PINC. More information about ST and TT speaking rates will be provided in §3.1.

2.1.2 Topics

As in the remaining corpora of the EPIC suite, specific topics of debates taking place at the European Parliament have been grouped into more general categories including agriculture and fisheries, economics and finance, employment, environment, health, justice, politics, procedure and formalities, science and technology, society and culture. There are topics that dominate the EP agenda, such as politics or economics and finance and those that are only occasionally discussed, e.g. science and technology, so an even distribution of topics across such a corpus is always difficult to obtain. Moreover, MEPs from different countries are not equally active in all debates. A perfect balance of topics is impossible, but it is still vital to be able to control the impact of the topic in those empirical investigations that require it. In the data selection process we paid particular attention to achieving a relatively even distribution of read and impromptu speeches across topics, although in the end it was not always possible. Also, the distribution of

3 Fluency parameters in the Polish Interpreting Corpus (PINC)

Table 1: Basic data about four subcorpora of PINC

		Number of speeches	Number of tokens	Average speech rate (wpm)
Polish source texts	Impromptu	117	20769	127
	Read	115	19399	126
English interpretations	Impromptu	117	21627	127
	Read	115	19656	131
English source texts	Impromptu	115	25715	178
	Read	115	28374	165
Polish interpretations	Impromptu	115	17496	121
	Read	115	19153	110

the two modes of delivery across topics in the two source language subcorpora, i.e. Polish and English, differs (Figure 1 and 2).

It transpires from Figure 1 and 2 that, in the PINC dataset, speeches regarding economics and finance are more often delivered impromptu by native English speakers at the EP, while the Polish MEPs read them slightly more often. Even more striking differences concern the speeches on politics, where most English speakers read texts out loud and the Polish ones predominantly speak impromptu.

2.1.3 Speakers

The ST-EN subcorpus contains speeches of 65 unique speakers (20 female and 45 male), while the ST-PL includes 57 unique speakers (11 female and 46 male). With 230 ST-EN speeches and 232 ST-PL speeches, this gives the average of 3.8 speeches per person (ranging from 1 to 19) in both subcorpora. Since PINC meta-data includes precise speaker identification, we will control for the uneven number of speeches in our analyses, whenever possible. We took extra care to exclude any non-native speakers of either language and to only include MEPs.

Interestingly, the majority of speeches delivered by Polish female MEPs were read out while the majority of male MEPs spoke impromptu. This was also true for English-speaking MEPs, although the differences are not as pronounced (Figure 3).

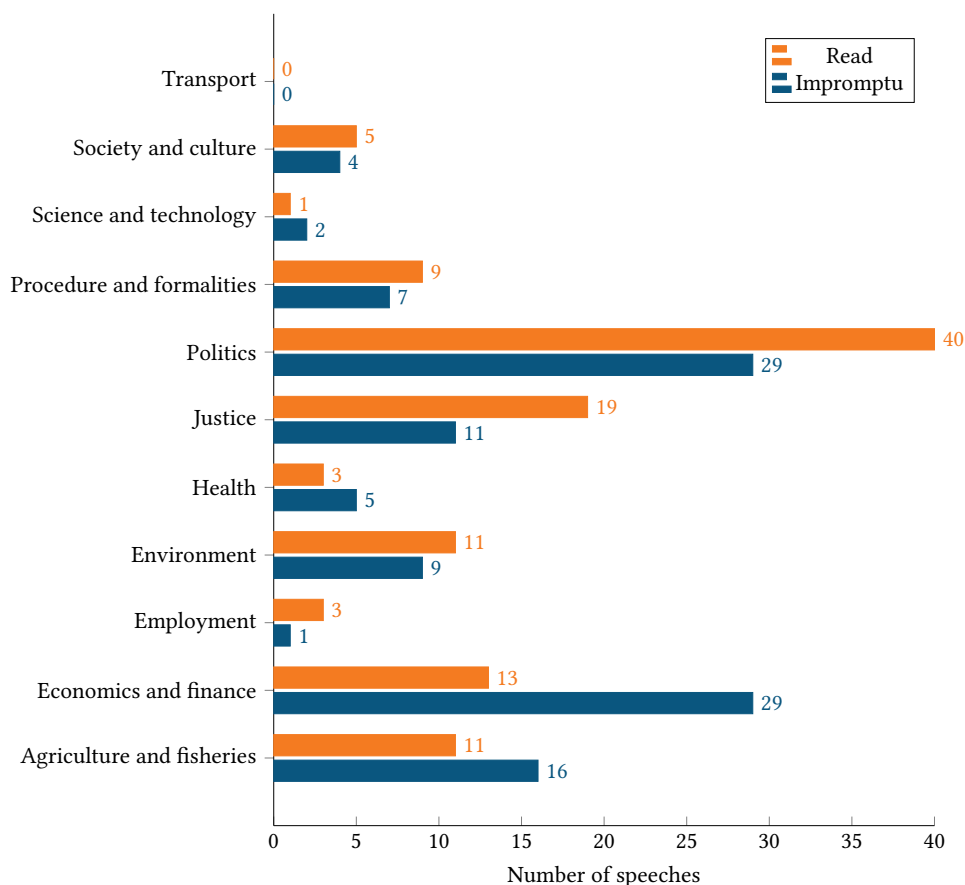


Figure 1: Topic coverage and mode of delivery in PINC subcorpora EN

2.1.4 Interpreters

Interpreters are key in any interpreting corpus. Professionals working during the European Parliament plenary sessions are carefully selected in a process designed to guarantee top quality interpreting services at the EU institutions. The usual problem with EP data, however, is that the only detail allowing us to distinguish between them is their voice. As most interpreting corpora are compiled by interpreting scholars with no expertise in speaker identification, interpreter identity in the corpora of the EPIC suite is frequently disregarded. Yet, controlling for individual variation is desired in many empirical studies, hence PINC does include precise metadata on interpreter identity. This will greatly enhance the control of the individual variation in further analyses.

3 Fluency parameters in the Polish Interpreting Corpus (PINC)

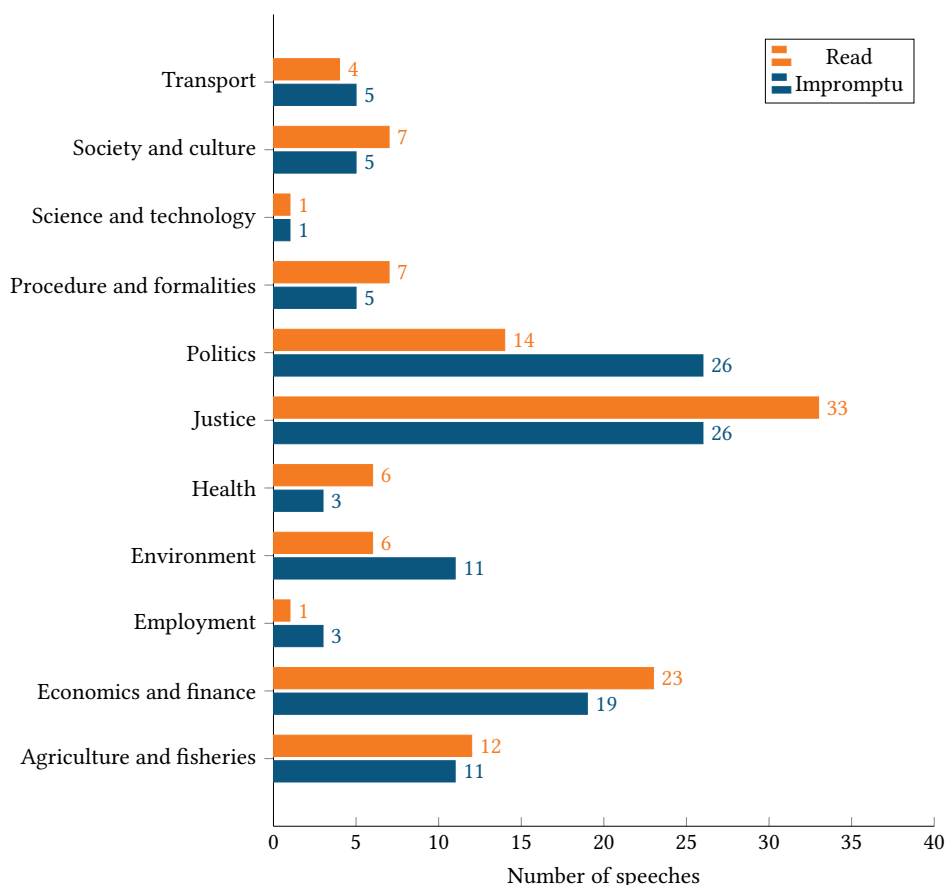


Figure 2: Topic coverage and mode of delivery in PINC subcorpora PL

Both interpreting subcorpora of PINC consist of slightly more texts interpreted by females. There are altogether 39 different interpreters in the PINC corpus, all of them Polish natives interpreting both into A (L1) and B language (L2). The TT-EN subcorpus contains texts interpreted by 21 different interpreters (10 female and 11 male) and the TT-PL subcorpus includes productions by 35 interpreters (23 female and 12 male). In most cases, interpreters interpreted both impromptu and read speeches in both directions (Figure 4 and 5), which makes it possible, for example, to investigate the same interpreter's interpreting output into different languages.

As not many interpreters in the European Parliament have Polish as a C language, interpretations from Polish are frequently provided as *retour* interpretations by interpreters from the Polish booth (with Polish as A and English as B), or

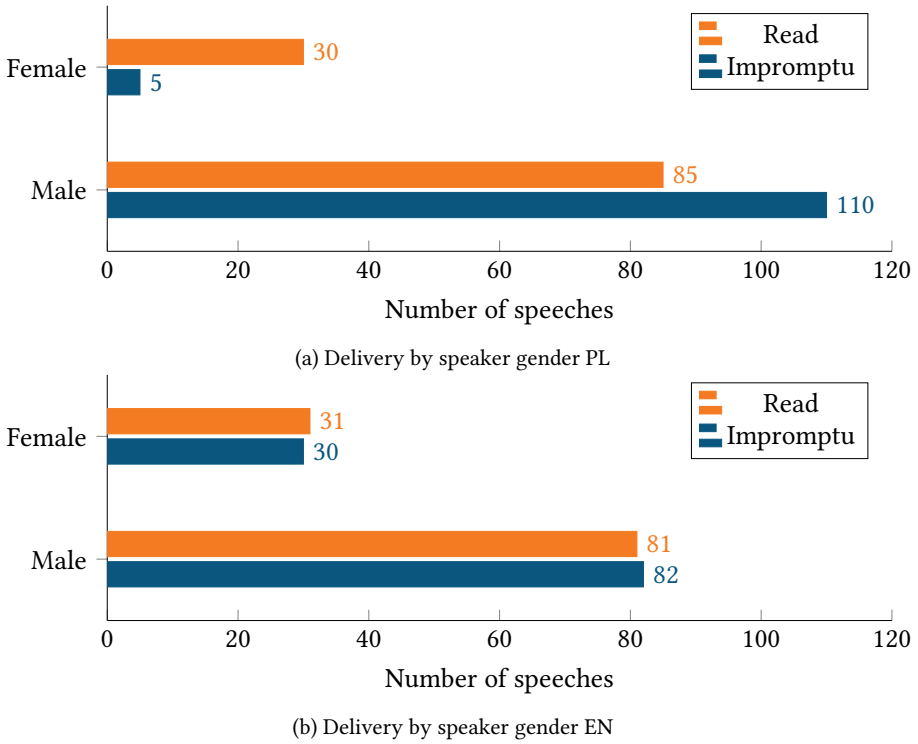


Figure 3: Speeches delivered by female and male MEPs in each mode

as relay interpretations when interpreters from other language booths use Polish-English retour as their pivot and the source input. We deliberately excluded any speeches interpreted via relay. As a result, the TT-EN subcorpus is a retour subcorpus and includes interpretations by the same interpreters who contributed to the TT-PL subcorpus. This offers an interesting opportunity for interlinguistic comparisons that are not between-groups, but within-group. This differentiates PINC from other corpora, which include either interpretations into A languages only or which do not strictly control for the language status (A or B) of the interpreters in specific subcorpora.

2.2 Design

2.2.1 Interpreter identification

Identifying interpreters may have presented the greatest technical challenge in building PINC so far. The Europarl website provides no information about the

3 Fluency parameters in the Polish Interpreting Corpus (PINC)

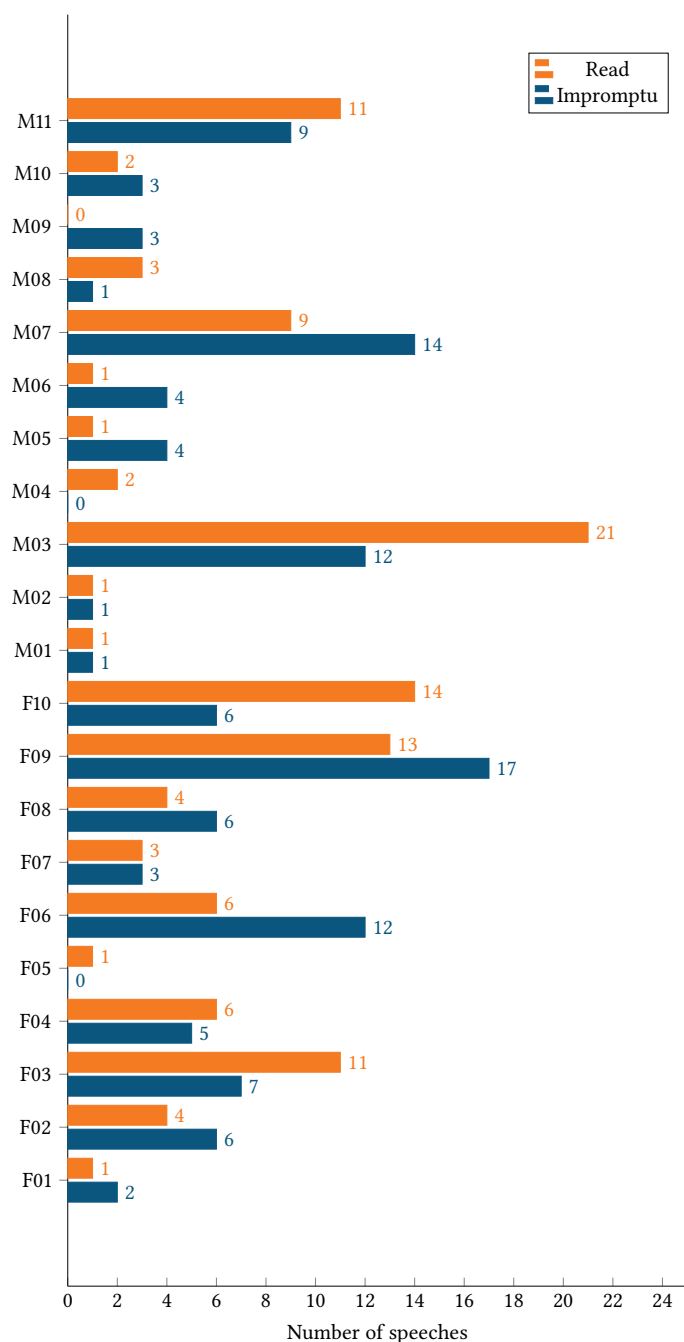


Figure 4: Read and impromptu speeches interpreted by individual interpreters PL-EN (codes starting with capital F indicate female interpreters, codes starting with M indicate male interpreters)

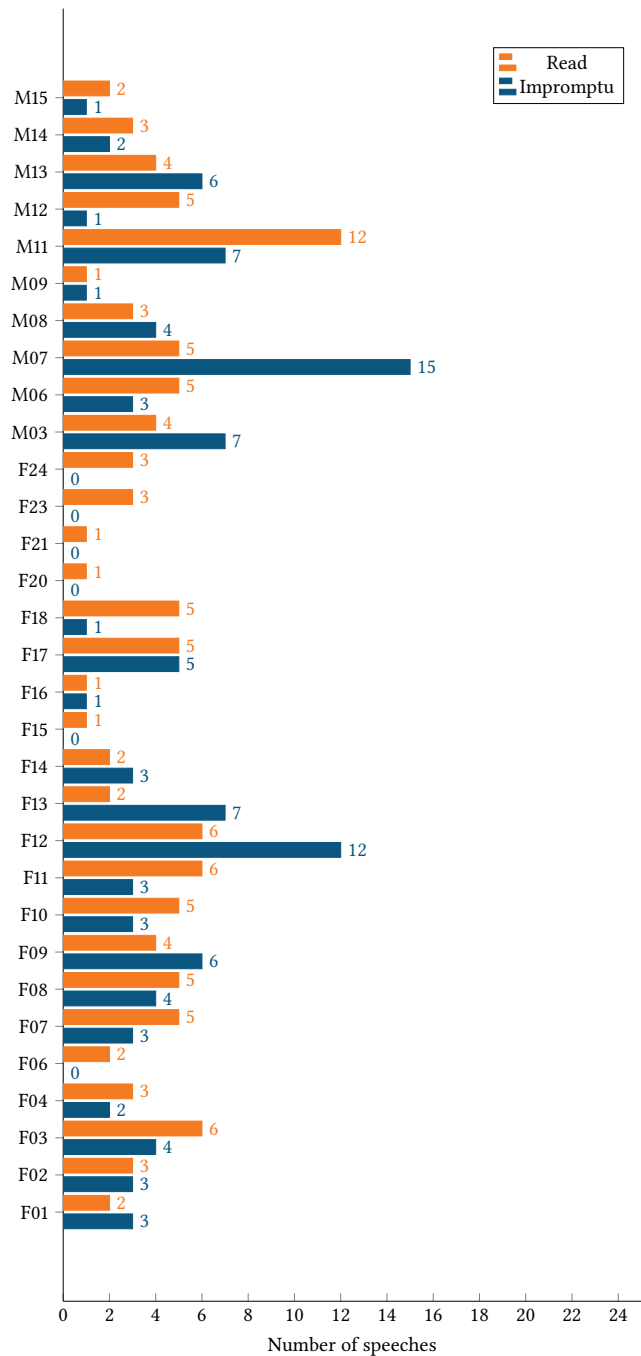


Figure 5: Read and impromptu speeches interpreted by individual interpreters EN-PL (codes starting with capital F indicate female interpreters, codes starting with M indicate male interpreters)

individual interpreters and, as opposed to the original speakers, they are not visually identifiable. In order to distinguish between the voices, we had to employ a three-stage procedure. In stage one, two human compilers who took part in collecting the corpus data (authors of this chapter) labelled each new interpreter in a spreadsheet. These were later proofed by another team member, especially where any doubts as to potential overlaps were expressed. In this manner, a pool of potential interpreter voices was identified. In stage two, this pool of identified voices was given for verification to an experienced conference interpreter who had worked with the interpreters included in the sample.

Independently of this strictly human-based procedure, in stage three, an automated attempt at interpreter identification was also made. It consisted in comparing the above pool of potential interpreter samples (enrolment data) to the recordings of all 476 interpretations (test data) within the Kaldi Speech Recognition Toolkit (Snyder et al. 2018) trained on large-scale, open-source corpora of human voices (development data). The method relies on computing a multi-dimensional vector representation of an audio segment, known as the x-vector. This vector is computed both for the enrollment data and for all the test data. Next, a Probabilistic Linear Discriminant Analysis algorithm is used to compute a matrix of distances between each file and speaker, thus providing an easy method of assigning the most likely candidate for each file. Interestingly, the comparison of the human-made and automatic judgments yielded very satisfying results as only around 15% of stage-one interpreter judgments have been misassigned. A detailed description of the interpreter voice identification procedure is provided in Koržinek (2020b).

2.2.2 Transcription

As in most of the corpora from the EPIC suite (Bernardini et al. 2018), the source text subcorpora in PINC are based on verbatim reports, i.e. transcripts of the audio/video files of speeches downloaded from the EP website. The EP website offers relatively accurate renditions that had to be manually corrected to a small extent only in order to facilitate speech-text alignment. Unfortunately, the texts of interpretations available on the EP website are actually written translations of the original verbatim reports and thus depart heavily from what was said by the interpreters. Therefore, in the case of target text subcorpora, we decided to use an automatic speech recognition system as input for later manual correction. We specifically used Google Cloud Speech as accessed through the WebMaus service (Kisler et al. 2017). To streamline the process of post-editing we set up a simple service based on the Corrector webApp (Koržinek 2019) consisting of a rich audio

player (based on wavesurfer.js audio editor and controllable from the keyboard) and a text field with basic change-tracking capabilities (Figure 6). Its cloud-based storage of results allowed for seamless cooperation between team members.

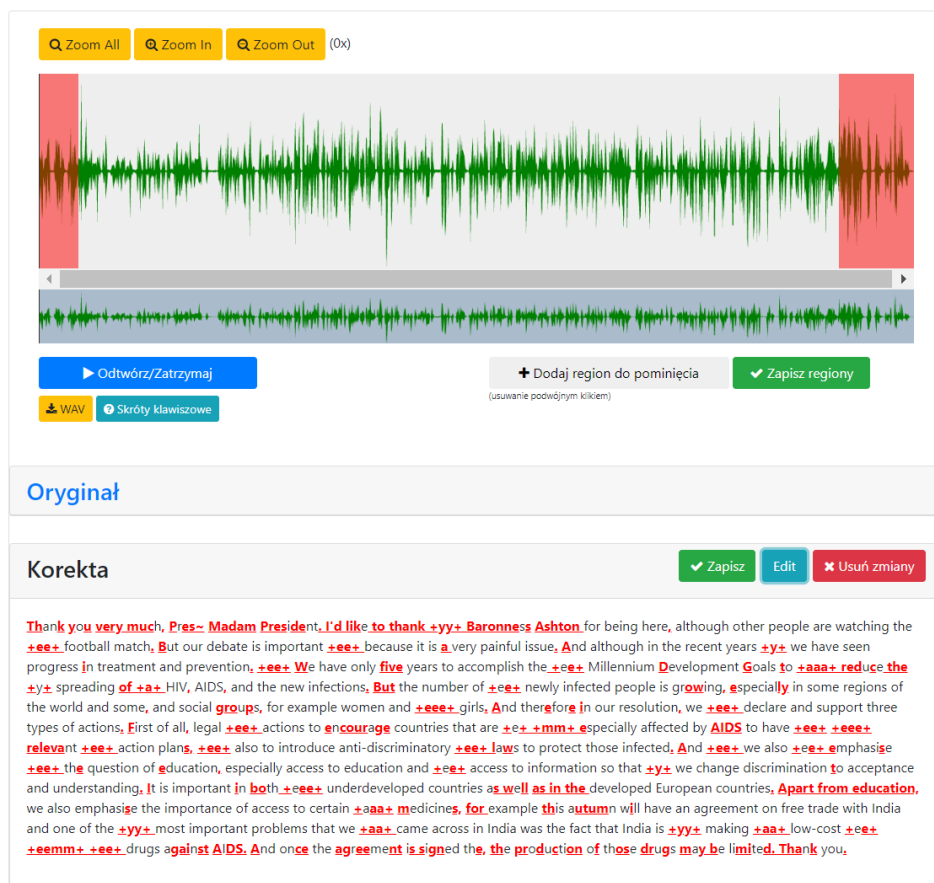


Figure 6: Corrector-webApp online environment

This application was used not only for correcting transcriptions but also for manual endpointing, that is, marking when the transcription starts and ends within the audio file (the pink areas in the waveform in Figure 6). Unfortunately, each audio recording begins and ends with a portion of speech that has a more dialogical and organisational character, such as the President giving the floor to a particular MEP whose speech is of primary interest in a given file. Thanks to endpointing, the alignment tools described below only utilised the audio that perfectly matched the transcription.

In terms of principles our transcription was largely based on guidelines used in EPIC and EPTIC (Bernardini et al. 2018: 27), altered in order to meet the needs of PINC. One such need was the automatic speech-text alignment described below; another was the planned ST-TT alignment on the word level. This required as accurate a marking of word-boundaries as possible, including unfinished, self-corrected and distorted words. In this respect we decided to introduce three special symbols: tildes <~> for truncated words, plus signs <+ +> marking the boundaries of filled pauses (the pluses were to surround an approximation of the actual sounds produced by the speaker, e.g. +ehm+) and square brackets <[]> to mark any external noises, such as applause, which could be picked up by the system and misinterpreted.

2.2.3 Speech-text alignment

While some corpora in the EPIC suite include only transcripts e.g. EPIC, others contain recordings that are time-aligned at various levels. Most language components of EPTIC are time-aligned with videos of the speeches at sentence-like utterance level (Ferraresi & Bernardini 2019: 132) using a system of subtitles integrated into NoSketchengine online platform (Rychlý 2007). EPICG includes timestamps at event level aligned in EXMARaLDA (Schmidt & Wörner 2009). PINC has been automatically time-aligned to audio files of the speeches/interpretations. The word-level alignment was then manually corrected in yet another instance of computer-human interaction employed for best possible results. This time the starting point was automatic segmentation and alignment performed in the Kaldi toolkit (Povey et al. 2011) and based on a Gaussian mixture-based acoustic model for which the endpointed transcriptions from the previous step were used as input along the audio recordings. Following that, two human aligners manually proofed and adjusted the output using the EMU-webApp (see Figure 7), which is an open-source browser-based labelling and correction tool that allows for a hassle-free cooperative annotation of audio files (Winkelmann & Raess 2014).

As a result, all words, pauses and disfluencies are orthographically transcribed, timestamped and available for analysis. A detailed description of the speech-text alignment in PINC is presented in Koržinek (2020a).

Further processing of the corpus (currently underway) involves pos tagging, text and video alignment, alignment of source texts and target texts on the utterance level and – most importantly – word level. This last alignment is especially crucial for the main objectives driving PINC creation. Apart from a plethora of

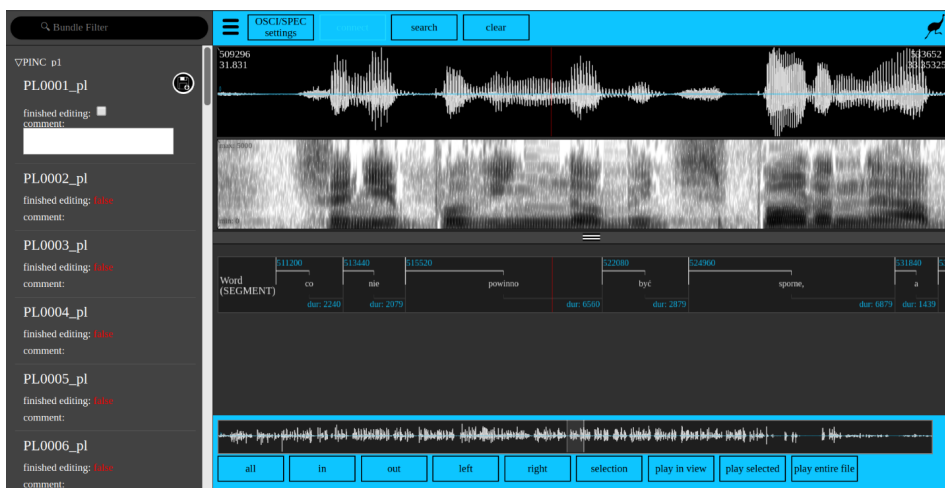


Figure 7: EMU-webApp online environment

corpus-driven research, PINC will first and foremost inform corpus-based studies on activation and inhibition as mechanisms of language control in interpreting. This is why precise timestamps are needed for specific words (cognates, homonyms, words with single and multiple translation equivalents), since we are interested, among other things, in the ear-voice span as a processing index of these words.

3 An example study: fluency parameters in interpreting

To show the potential of PINC, we present an example of a study that looks into interpreting fluency parameters, such as speaking rate and pauses. We compared source texts and their interpretations on a number of delivery parameters and tried to identify which factors modulate these parameters in interpreters' outputs. We also wanted to find out if interpreters speed up and compress their target text production when dealing with higher source text delivery rates. Thanks to interpreter voice identification in PINC metadata, we could explore individual differences and control for these differences in our analysis. We conducted the study on the Polish-English subcorpus, so the interpreting examined is performed into the interpreters' B language, i.e. the more demanding interpreting direction (Chang 2005; see also a review in Chmiel 2016).

3.1 Interpreting speed and its modulating factors

Speed of delivery is considered one of the most important input variables in interpreting, which has been shown to affect the quality of interpreting (Riccardi 2015), including omissions (Barghout et al. 2015) or the occurrence of filled pauses (Plevoets & Defrancq 2016). While the majority of studies focus on source text speed as an important factor that influences numerous aspects of interpreters' output, few studies have specifically focused on various factors that affect the target text speed. For instance, Han (2015) found that speech rate in interpreting has a strong correlation with perceived fluency. Below, we use PINC data to see what makes interpreters speed up their production. First, however, we analyse the corpus to compare ST and TT speeds on a number of measures and discuss our results in the context of other available data on comparable corpora.

As mentioned above, the average speaking speed in our Polish-English sub-corpus was 126 wpm (SD=15, range: 88–166) for ST and 129 wpm (SD=18, range: 77–178) for TT. These values are considered as low speed of delivery by EPIC standards (Monti et al. 2005) and are lower than those reported for EPICG (158 wpm for ST and 142 wpm for TT, Collard & Defrancq 2019). The ST speaking speed is also lower than 154 wpm from EPIC reported by Russo (2018) while the TT speed is comparable with the relevant data from the same study (130 wpm).

As languages may differ in word length, some researchers (Riccardi 2015; Seiber 2017; Tissi 2000) pinpoint that speaking speed may also be measured in syllables per minute. When measured this way, the PINC source texts are characterised by a significantly higher speaking rate ($M=286$ spm, $SD=35$) than target texts ($M=199$ spm, $SD=27$), $t(433)=30.26$, $p<.001$, which results from the fact that Polish words are on average longer (2.27 syllables per word in our corpus) than English ones (1.55 syllables per word in our corpus).

Another important measure of the speed of oral delivery is the articulation rate understood as the average speed of utterance without pauses (Christodoulides 2013; Riccardi 2015). The PINC source texts have a higher articulation rate (measured in syllables per minute, $M=330$ spm, $SD=36$) than target texts ($M=248$ spm, $SD=24$), $t(404)=28.48$, $p<.001$.

We also measured the compression rate understood, following Russo (2018), as a relative difference in speech length, expressed in percent and measured according to the following formula: $(\text{total ST words} - \text{total TT words}) * 100 / \text{total ST words}$. If the compression rate is 0, the target text equals the source text in length. If it has negative value, the target text is compressed. If the value is positive, the target text is longer than the original. The mean compression rate for the whole subcorpus is 3.6%, which means that the interpretations are slightly longer than

the originals. However, there is much variation among individual interpreters, and we can visualise that thanks to exact identification of interpreter voices in the corpus (Figure 8).

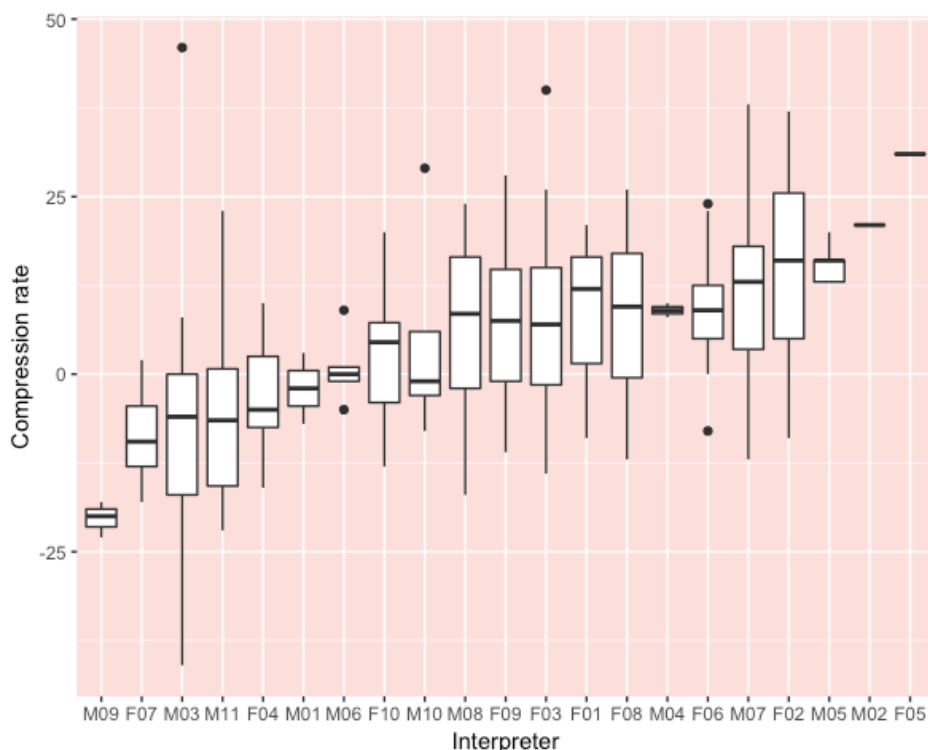


Figure 8: Individual variation of compression rates

There are six interpreters who consistently compress the source text while the majority of interpreters produce longer interpretations than originals, which is quite surprising and at variance with Russo (2018) but might be triggered by two factors. First, PINC source texts are slower than those analysed by Russo: interpreters might not feel compelled to synthesise if there are no demanding temporal constraints. Second, this analysis pertains to interpretations into B language only and these, as such, might differ in production characteristics from interpretations into A language, for example in terms of opting for more descriptive formulations where precise one-to-one equivalents are not easily retrievable from the mental lexicon. Further comparisons are needed, and they will be possible when PINC, as planned, is extended to include a smaller subcorpus of the

same language combination (PL-EN) with interpretations performed into A language.

In order to see whether interpreters speed up their delivery and compress more when processing a fast source text, we fitted three regression models. The data show that source texts with higher word per minute values lead to interpretations with higher word per minute values ($b=.49$, $SD=.07$, $t=6.96$, $p<.001$, Figure 9), with higher articulation rates ($b=.54$, $SD=.10$, $t=5.35$, $p<.001$, Figure 10) and higher compression rates ($b=-.40$, $SD=.05$, $t=-7.19$, $p<.001$, Figure 11).

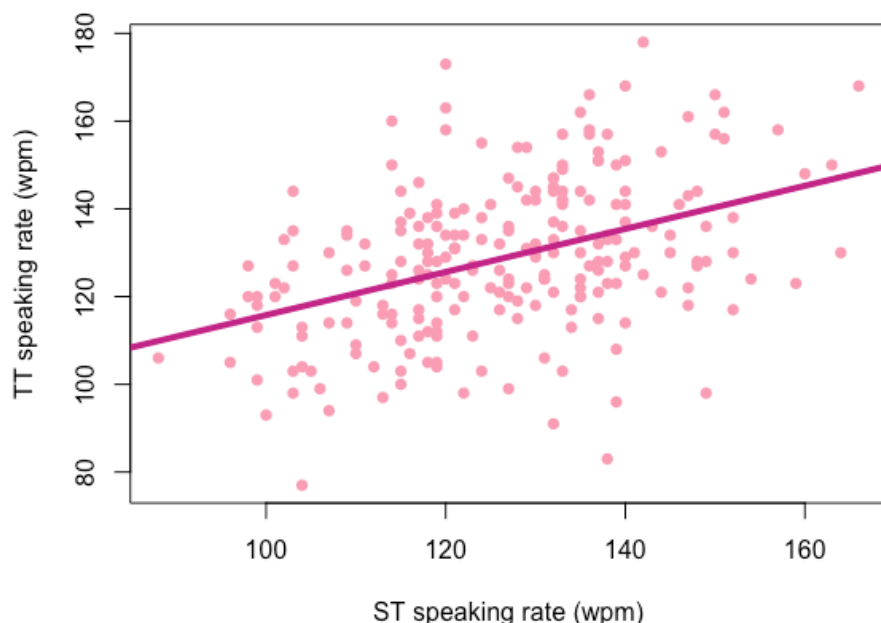


Figure 9: Mean source text speed (words per minute) plotted against target text speed (words per minute)

The results of our analysis regarding how interpreters modulate their output as a result of the source text speed are in line with those by Russo (2018), who also found that faster source texts lead to faster target texts and greater compression, and with those by Gerver (1969) and Barghout et al. (2015), who identified a similar relation between ST speed and TT compression. Slower speaking by interpreters as compared to source text speakers was previously confirmed by

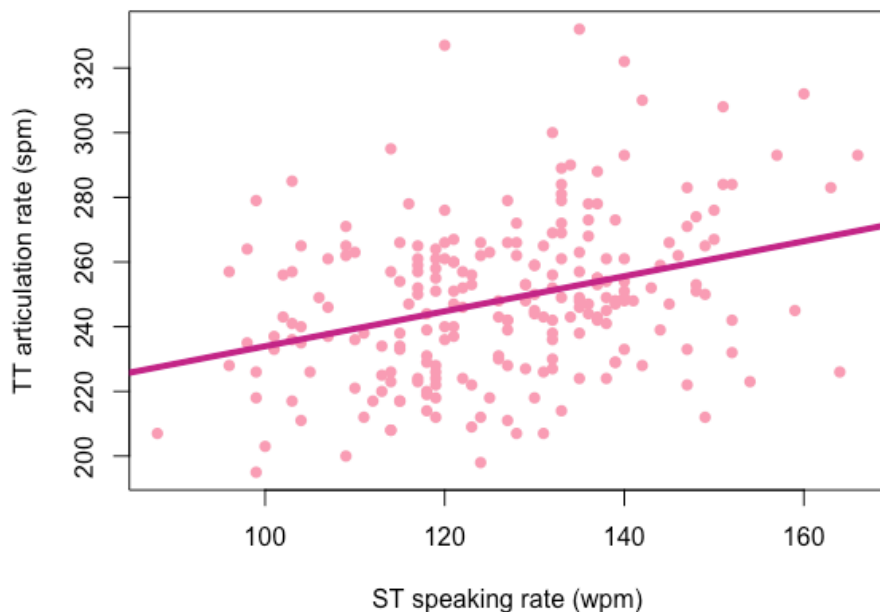


Figure 10: Mean source text speed (words per minute) plotted against target text articulation rate (syllables per minute)

Russo (2018) and Christodoulides (2013). This might be explained by the fact that – due to compression – interpreters speak less and thus can slow down.

Taken together, these results show an established pattern of the source text speed affecting interpreters' production in terms of speed and compression. The novelty of PINC analysis is the option to better capture individual differences by using specific metadata regarding interpreter voice identification. For instance, in an additional analysis, we used the range of speaking rate of each speaker and interpreter as a variable of speaking rate variability. We excluded from this analysis those speakers and interpreters who only contributed one speech to the data set (as their variability was 0), which left us with data from 42 speakers and 20 interpreters to analyse. It turned out that the individual speaking rate variability of interpreters was much higher ($M=40$ wpm) than that of speakers ($M=19.95$ wpm), $t(23.8)=4.60$, $p=.0001$. This confirms the results obtained by Christodoulides (2013) on a much smaller corpus based on EP data.

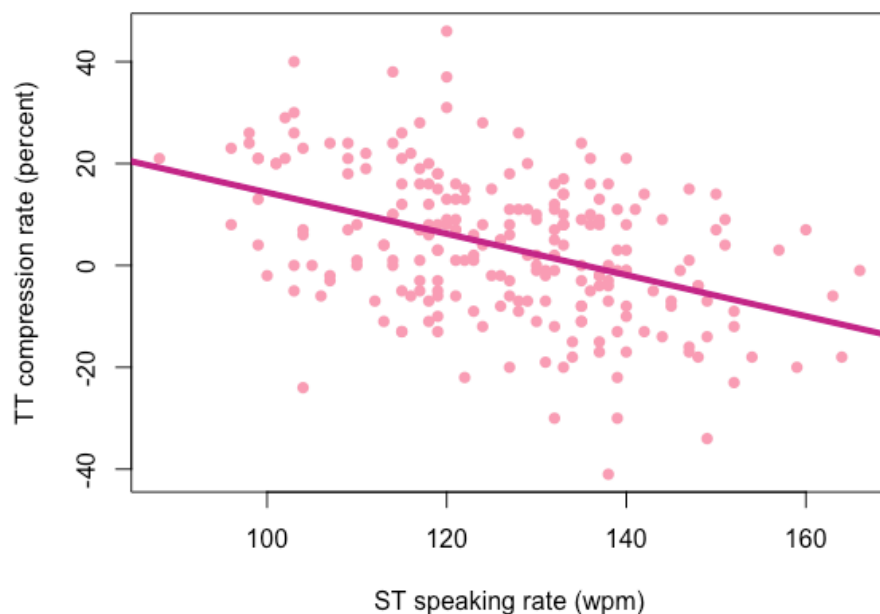


Figure 11: Mean source text speed (words per minute) plotted against compression rate (percent)

3.2 Comparing other delivery parameters in source and target texts

Many studies compare fluency parameters of source texts and their interpretations to shed more light on the difference between non-mediated and interpreter-mediated texts (Ahrens 2005; Cecot 2001; Pöchhacker 1995; Tissi 2000; Wang & Li 2014). The emerging pattern of data resulting from these experimental studies is that the pausing pattern specific to interpreter-mediated texts includes fewer but longer pauses. Our analysis makes a contribution to the corpus data on ST-TT comparison. The novelty of our findings is that they are based solely on the subcorpus of retour interpretations (i.e. interpretations into the interpreter's B language). Below we compare source and target texts in the Polish-English subcorpus of PINC on a range of parameters other than those related to speed analysed above and pertaining mainly to pauses. We later compare our results to other findings based on other corpora.

The comparisons of PINC source and target texts are presented in Table 2. Duration was calculated from the onset of the first spoken word to the ending of the last word of each speech. Thus, it did not include silence periods before and after the utterance. This is quite important to remember since the TT duration does not include the initial ear-voice span (EVS) and as such does not capture the dynamics of processing involved in interpretation. A detailed analysis of EVS will be the focus of another study. Further in Table 2, there are four parameters pertaining to silent and filled pauses – reflecting their number (normalised per minute of speech) and mean length. Various thresholds are used in the literature to identify silent pauses, ranging from 200 ms (Chmiel et al. 2017; Collard & Defrancq 2019) to 300 ms (Wang & Li 2014). We identified a silent pause as a period of silence longer than 250 ms, in line with the majority of studies (Cecot 2001; Han et al. 2020; Mead 2005; Pradas Macías 2006; Tissi 2000). A filled pause was identified as anything marked in transcription as +yyy+ or +eee+ or anything else between two plus signs. We applied no cut-off point for a filled pause following Plevoets & Defrancq (2016). A run was defined as a segment of speech uninterrupted by silent pauses, as applied by Han et al. (2020). Finally, speech proportion was calculated as a ratio of articulation time (i.e. not including pauses) to speech duration (Lee 1999).

Table 2: ST and TT delivery parameters compared

Parameter	ST mean	TT mean	t	p
Duration	1 min 22 s	1 min 24 s	-.28	=.78
Number of silent pauses per minute	12.06	10.36	5.51	p<.001*
Mean length of silent pauses (in ms)	487	626	-6.86	p<.001*
Number of filled pauses per minute	2.84	7.37	-11.10	p<.001*
Mean length of filled pauses (in ms)	613	722	-3.99	p<.001*
Mean length of runs (in syllables)	26.21	20.20	2.56	p=.011*
Speech proportion	0.87	0.80	11.38	p<.001*

Differences in all the parameters apart from duration turned out to be statistically significant. The comparison shows a familiar pattern: interpretations include fewer but longer silent pauses, which is in line with a study involving students by Tissi (2000), a small-scale study of A to B simultaneous interpreting involving professionals and trainees by Wang & Li (2014) and other studies (Ahrens 2005; Christodoulides 2013; Collard & Defrancq 2019; Lee 1999; Pöschhacker 1995).

It is interesting to see that silent pauses in PINC are much shorter and less numerous than in a comparable corpus (EPICG) involving interpretations from the European Parliament and featuring different language pairs analysed by Collard & Defrancq (2019). In that study, there are almost 23 silent pauses in one minute of ST and 19 silent pauses in one minute of TT. The mean length is 10280 ms and 10580 ms for ST and TT, respectively. It seems that, as compared to PINC, silent pauses in EPICG are approximately twice as long and twice as numerous in all texts. We might speculate that this discrepancy is due to differences in speaking rates, which are much higher in EPICG. A comparison of PINC and EPICG data for filled pauses is also interesting. Despite differences in speaking rates, the numbers of filled pauses per minute match almost exactly across both corpora: 2.61 in EPICG and 2.84 in PINC for source texts and 7.52 in EPICG and 7.37 in PINC for target texts. Unfortunately, Collard and Defrancq do not include data for the mean length of filled pauses. A potential explanation for these results might be the different nature of both corpora. EPICG includes, to the best of our knowledge, only interpretations into the A language, while the subcorpus of PINC under analysis includes *retour* interpretations only (i.e. into the B language). Since production in one's B language is more difficult than in one's A language and since filled pauses, according to Setton (1999), reflect cognitive load related to formulation, interpretations into B should include a greater number of filled pauses than interpretations into A. The reason PINC and EPICG match on this measure might be because the number of filled pauses in PINC is offset by its lower ST and TT speaking rate. This explanation is tentative and the prediction on the higher number of filled pauses present in *retour* interpretations as compared to interpretations into A will be tested on the Polish-English language pair once PINC is extended to include a subcorpus of PL-EN interpretations made into the interpreters' A language.

To the best of our knowledge, no study to date has compared the mean length of runs of source texts and target texts in a corpus study. Our data show that interpretations include shorter runs, or uninterrupted flows of utterance, than source texts. This might mean that interpreters work in shorter spurts and fragment their output due to processing constraints. Additionally, the speech pro-

portion data are in line with Lee's (1999) results, showing that interpreters speak for a smaller proportion of time than speakers and use pauses for information processing.

In their study of perceived fluency of interpreting, Han et al. (2020) identified the following criteria as strongly associated with higher fluency: mean length of runs, mean length of silent pauses, phonation time ratio (which is equivalent to speech proportion in the present study) and speech rate. All these criteria have lower values for interpreting than for the source texts in PINC. Although Han et al.'s (2020) data pertain to consecutive interpreting, we might tentatively assume that interpretations in our corpus could be perceived as less fluent than the source texts, although such a conjecture surely requires empirical verification.

3.3 Silent and filled pauses in interpreting and their modulating factors

Fluent delivery is an important criterion in interpreting (Pradas Macías 2006; Rennert 2010) and pauses are generally considered as an important element of fluency (Mead 2000). Pöschhacker (2004) considers silent and filled pauses as part of the disfluency phenomenon in interpreting related to the limited scope of planning involved in this type of oral production. Silent pauses are associated with problems with ST comprehension, lexical search for translation equivalents and production difficulties (Bartłomiejczyk 2006, Piccaluga et al. 2005, Tóth 2011). Although interpreters tend to follow the general pattern of pauses applied by the speaker, there are modifications to that pattern due to difficulties in processing (Cecot 2001; Goldman-Eisler 1972). Filled pauses can be interpreted as an indirect index of cognitive load and, similarly to silent pauses, can reflect processing difficulties. According to Setton (1999), while long silent pauses indicate high attention to input, long filled pauses reflect attention to formulation including speech planning and lexical access. Both silent and filled pauses turned out to be longer in interpretations than in source texts, thus testifying to extreme speech production conditions in interpreting.

We fitted a series of mixed effects linear models to evaluate the impact of ST fluency parameters on pauses in the target texts. Since PINC metadata include exact identification of interpreter voices, we could include interpreters as a random factor in each model. Fixed factors reflected the source text delivery characteristics, such as speed, number of silent pauses per minute, mean length of silent pauses, compression rate and delivery mode (whether the source text was read or delivered impromptu). We could not include the number and mean length of

filled pauses due to the violation of the collinearity principle – both of these measures correlated moderately with the source text speaking rate. We used sliding contrasts for delivery mode and treatment contrasts for the remaining fixed factors. P values were obtained through Satterthwaite approximations. The number of silent pauses per minute in the interpretation was influenced by the ST speed ($b=-.02$, $SE=.01$, $t=-3.03$, $p=.002$), mode of delivery ($b=-1.04$, $SE=.44$, $t=-2.36$, $p=.02$) and the number of silent pauses per minute in the source text ($b=.15$, $SE=.07$, $t=2.23$, $p=.03$). The faster the ST, the lower the number of silent pauses in TT. There are more silent pauses in interpreting read-out speeches than impromptu speeches and the number increases when there are more silent pauses in ST. This last association is in line with Collard & Defrancq (2019). It seems that interpreters pause less when dealing with faster and read-out source texts, but they pause more when the speakers pause more. Interestingly, the mean length of silent pauses in TT was not modulated by any factors, which is at variance with Collard and Defrancq's study, where source text speaking rate did modulate the length but not the number of silent pauses in interpreting.

The number of filled pauses in TT was modulated by the mode of delivery ($b=1.24$, $SE=.58$, $t=2.12$, $p=.03$) and the compression rate ($b=-9.67$, $SE=2.39$, $t=-4.05$, $p<.001$). The data show that as compression increases and the interpretation becomes shorter, the number of filled pauses increases. There were also more filled pauses in read out speeches as compared to the impromptu ones. As postulated earlier, this might be related to the cognitive load triggered by increased reformulation involved in producing more compressed and structurally less complex target texts. The mean length of filled pauses in TT was modulated by the ST speed ($b=-1.94$, $SE=.56$, $t=-3.45$, $p<.001$) and the compression rate ($b=-673.42$, $SE=150.58$, $t=-4.47$, $p<.001$). These results show that the faster and the less compressed the source text, the lower the mean length of filled pauses in the interpretation. None of the predictors associated with the number of filled pauses in this study match those in EPICG (Collard & Defrancq 2019). In that study, ST speed influenced the number of filled pauses while in ours it influenced the length of filled pauses. More research is needed to elucidate the phenomenon of filled pauses in interpreting.

Taken together, our data on factors modulating pauses are partially in line with Setton's (1999) general idea of the relationship between silent pauses and the focus on the ST input and between filled pauses and the focus on formulation. Only filled pauses were modulated by the compression rate: they became longer and more numerous as interpreters struggled to provide a more compressed, i.e. more reformulated, version of the target text. Mode of ST delivery influenced

the number of both silent and filled pauses. They were more numerous in interpretations of read-out speeches. One may assume that silent pauses helped interpreters' comprehension of these speeches that are usually lexically denser and structurally more complex. Filled pauses, on the other hand, aided formulation, which was also more demanding as compared to impromptu speeches that are usually more similar in structural complexity to oral production involved in interpreting. However, these conjectures require further empirical support. The study by Wang & Li (2014) constitutes an interesting attempt at providing detailed explanations of various categories of pauses thanks to a combination of experimental data with retrospective protocols. Alas, no differentiation between motivations for silent and filled pauses is made. This definitely is a promising research avenue worth pursuing in the future.

4 Conclusions

PINC offers excellent research material that is well balanced, considering the external constraints. Issues of justice and politics predominate the topics of speeches, while as far as gender distribution is concerned, the majority of speakers are male and the majority of interpreters are female. This mirrors the European Parliament reality – male MEPs still dominate the chamber while interpreting is unceasingly a profession dominated by females (which is also true for experimental studies as gender balance is difficult to gain when recruiting study participants). The PINC creation workflow offers new tools and automation opportunities for future corpus developers.

Thanks to using similar categories of metadata (topics, mode of delivery), PINC will be easily comparable to other corpora from the EPIC suite, which should facilitate studies that involve various language combinations to control for language-pair-specific factors. Interestingly and due to the language regime and language profiles of interpreters in the European Parliament, PINC includes a strictly controlled Polish-English subcorpus of *retour* interpreting, an added value as compared to other existing corpora. In the future, it will also include a smaller subcorpus of Polish-English interpretations into A language. This offers a lot of potential for various novel comparisons in corpus-driven studies. We can compare interpretations by the same interpreters working into A (EN-PL) and B (PL-EN). We can also compare interpretations in the same direction (PL-EN) by two different groups of interpreters – native speakers of English and interpreters with English as their B language.

Our initial exploratory corpus-driven study shows how important it is to apply various variables since not all of them are sensitive enough to capture differences.

Our ST-PL and TT-EN corpora differed in speaking rate measured in syllables per minute but not in words per minute. Interestingly, the mean compression rate was slightly positive, meaning that target texts were actually longer than source texts. However, a detailed analysis of individual differences showed compression as an interpreter-specific feature. We found that interpreters speed up and compress their delivery more when the source text is delivered faster, showing an expected pattern of results in line with previous studies. Our source and target texts differed also on a range of other fluency criteria, such as number and mean length of silent and filled pauses. We also applied another measure of fluency – mean length of runs (i.e. utterances uninterrupted by pauses) – and found interpreters to produce more fragmented output due to processing constraints. Our findings show that interpreters produce more silent and filled pauses when interpreting a read-out text. More numerous silent pauses in the source text also increase the number of such pauses in the target text. Additionally, the number and length of filled pauses increase with increased compression rate, which seems to suggest that filled pauses could be a good index of production problems.

PINC has been created mainly to expand our knowledge about language control mechanisms (activation and inhibition) in interpreters on the basis of naturalistic data and to serve as a source of stimuli for future experimental studies. However, we hope that PINC, with its intended open access format, rich annotation and built-in interpreter identification will also help interpreting scholars find answers to many interesting corpus-driven research questions.

Acknowledgements

This research was supported by the UMO-2018/30/E/HS2/00035 (SONATA BIS 8) research grant funded by the National Science Centre, Poland (2019-2024).

References

- Ahrens, Barbara. 2005. Prosodic phenomena in simultaneous interpreting: A conceptual approach and its practical application. *Interpreting* 7(1). 51–76. DOI: 10.1075/intp.7.1.04ahr.
- Barghout, Alma, Lucía Ruiz Rosendo & Mónica Varela García. 2015. The influence of speed on omissions in simultaneous interpretation: An experimental study. *Babel* 61(3). 305–334. DOI: 10.1075/babel.61.3.01bar.
- Bartłomiejczyk, Magdalena. 2006. Strategies of simultaneous interpreting and directionality. *Interpreting* 8(2). 149–174. DOI: 10.1075/intp.8.2.03bar.

- Bartłomiejczyk, Magdalena. 2016. *Face threats in interpreting: A pragmatic study of plenary debates in the European Parliament*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.
- Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard & Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, 21–42. Singapore: Springer. DOI: 10.1007/978-981-10-6199-8_2.
- Cecot, Michela. 2001. Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters' performances. *The Interpreters' newsletter* 11. 63–85.
- Chang, Chia-chien. 2005. *Directionality in Chinese/English simultaneous interpreting: Impact on performance and strategy use*. Austin: University of Texas. (Doctoral dissertation).
- Chmiel, Agnieszka. 2012. Pamięć operacyjna tłumaczy konferencyjnych mierzona metodą RSPAN. In Maria Piotrowska (ed.), *Kompetencje tłumacza*, 137–154. Kraków: Tertium.
- Chmiel, Agnieszka. 2016. Directionality and context effects in word translation tasks performed by conference interpreters. *Poznań Studies in Contemporary Linguistics* 52(2). 269–295. DOI: 10.1515/psicl-2016-0010.
- Chmiel, Agnieszka. 2018. Meaning and words in the conference interpreter's mind: Effects of interpreter training and experience in a semantic priming study. *Translation, Cognition & Behavior* 1(1). 21–41. DOI: 10.1075/tcb.00002.chm.
- Chmiel, Agnieszka, Agnieszka Szarkowska, Danijel Koržinek, Agnieszka Lijewska, Łukasz Dutka, Łukasz Brocki & Krzysztof Marasek. 2017. Ear–voice span and pauses in intra- and interlingual respeaking: An exploratory study into temporal aspects of the respeaking process. *Applied Psycholinguistics* 38(5). 1201–1227. DOI: 10.1017/S0142716417000108.
- Christodoulides, George. 2013. Prosodic features of simultaneous interpreting. In Piet Mertens & Anne-Catherine Simon (eds.), *Proceedings of Prosody-Discourse Interface Conference 2013 (IDP-2013)*, 33–37. Leuven: University of Leuven.
- Collard, Camille & Bart Defrancq. 2019. Disfluencies in simultaneous interpreting a corpus-based study with special reference to sex. In Lore Vandevoorde, Joke Daems & Bart Defrancq (eds.), *New empirical perspectives on translation and interpreting*, 264–299. New York: Routledge. DOI: 10.4324/9780429030376-12.
- Defrancq, Bart, Koen Plevvoets & Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to*

3 Fluency parameters in the Polish Interpreting Corpus (PINC)

- discourse and translation studies*, 195–222. Cham: Springer. DOI: 10.1007/978-3-319-17948-3_9.
- Dumara, Barbara. 2015. *How can interpreting corpora extend our knowledge on intrusive "we" in SI?* Poster presented at Corpus-based Interpreting Studies: The State of the Art. First Forlì International Workshop. Forlì, Italy.
- Ferraresi, Adriano & Silvia Bernardini. 2019. Building EPTIC: A many-sided, multi-purpose corpus of EU parliament proceedings. In Irene Doval & M. Teresa Sánchez Nieto (eds.), *Parallel corpora for contrastive and translation studies: New resources and applications*, 123–139. Amsterdam/Philadelphia: John Benjamins.
- Gerver, David. 1969. The effects of source language presentation rate on the performance of simultaneous conference interpreters. In *Proceedings of the Second Louisville Conference on rate and/or frequency-controlled speech*, 162–184. Louisville: Center for Rate-Controlled Recordings, University of Louisville.
- Goldman-Eisler, Frieda. 1972. Segmentation of input in simultaneous translation. *Journal of Psycholinguistic Research* 1(2). 127–140.
- Han, Chao. 2015. (Para)linguistic correlates of perceived fluency in English-to-Chinese simultaneous interpretation. *International Journal of Comparative Literature and Translation Studies* 3(4). 32–37. DOI: 10.7575/aiaa.ijclts.v.3n.4p.32.
- Han, Chao, Sijia Chen, Rongbo Fu & Qin Fan. 2020. Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interpreting* 22(2). 211–237. DOI: 10.1075/intp.00040.han.
- Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu. (Doctoral dissertation). <https://hdl.handle.net/10593/2425>.
- Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347. DOI: 10.1016/j.csl.2017.01.005.
- Koržinek, Danijel. 2019. *Corrector*. <https://github.com/danijel3/Corrector>.
- Koržinek, Danijel. 2020a. *Automating word segmentation*. PINC Project. <https://pincproject2020.wordpress.com/2020/04/08/automating-word-segmentation/>.
- Koržinek, Danijel. 2020b. *Speaker identification*. PINC Project. <https://pincproject2020.wordpress.com/2020/04/28/speaker-identification/>.
- Lee, Tae-Hyung. 1999. Speech proportion and accuracy in simultaneous interpretation from English into Korean. *Meta* 44(2). 260–267. DOI: 10.7202/003443ar.
- Mead, Peter. 2000. Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter* 10(200). 89–102.

- Mead, Peter. 2005. Directionality and fluency: An experimental study of pausing in consecutive interpretation into English and Italian. *Communication and Cognition. Monographies* 38(1-2). 127–146.
- Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli & Mariachiara Russo. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta* 50(4). 141–158. DOI: 10.7202/019850ar.
- Piccaluga, Myriam, Jean-Luc Nespoulous & Bernard Harmegnies. 2005. Disfluencies as a window on cognitive processing. An analysis of silent pauses in simultaneous interpreting. In Jean Véronis & Estelle Campione (eds.), *Disfluency in spontaneous speech*, 151–155. ISCA.
- Plevoets, Koen & Bart Defrancq. 2016. The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies* 11(2). 202–224. DOI: 10.1075/tis.11.2.04ple.
- Pöhhacker, Franz. 1995. Slips and shifts in simultaneous interpreting. In Jorma Tammola (ed.), *Topics in interpreting research*, 73–90. Turku: University of Turku, Centre for Translation & Interpreting.
- Pöhhacker, Franz. 2004. *Introducing interpreting studies*. London: Routledge.
- Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer & Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 1–4. Waikoloa: IEEE Signal Processing Society.
- Pradas Macías, Macarena. 2006. Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency. *Interpreting* 8(1). 25–43. DOI: 10.1075/intp.8.1.03pra.
- Rennert, Sylvi. 2010. The impact of fluency on the subjective assessment of interpreting quality. *The Interpreters' Newsletter* 15. 101–115.
- Riccardi, Alessandra. 2015. Speech rate. In Franz Pöhhacker (ed.), *Routledge encyclopedia of interpreting studies*, 1st edn., 397–399. London: Routledge.
- Russo, Mariachiara. 2018. Speaking patterns and gender in the European Parliament Interpreting Corpus: A quantitative study as a premise for qualitative investigations. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, 115–131. Singapore: Springer.
- Rychlý, Pavel. 2007. Manatee/Bonito: A Modular Corpus Manager. In Petr Sojka & Aleš Horák (eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, 65–70. Brno: Masarykova Univerzita.

- Schmidt, Thomas & Kai Wörner. 2009. EXMARaLDA: Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19(4). 565–582.
- Seeber, Kilian G. 2017. Interpreting at the European Institutions: Faster, higher, stronger. *CLINA: Revista Interdisciplinaria de Traducción, Interpretación y Comunicación Intercultural* 3(2). 73–90. DOI: 10.14201/clina2017327390.
- Setton, Robin. 1999. *Simultaneous interpretation: A cognitive-pragmatic analysis* (Benjamins Translation Library 28). Amsterdam: John Benjamins. DOI: 10.1075/btl.28.
- Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43(4). 486–493. DOI: 10.7202/004136ar.
- Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey & Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. Calgary, AB: IEEE. DOI: 10.1109/ICASSP.2018.8461375.
- Tissi, Benedetta. 2000. Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis. *The Interpreters' Newsletter* 10. 103–127.
- Tóth, Andrea. 2011. Speech disfluencies in simultaneous interpreting: A mirror on cognitive processes. *SKASE Journal of Translation and Interpretation* 5(2). 23–31.
- Wang, Binhua & Tao Li. 2014. An empirical study of pauses in Chinese-English simultaneous interpreting. *Perspectives* 23(1). 124–142. DOI: 10.1080/0907676X.2014.948885.
- Winkelmann, Raphael & Georg Raess. 2014. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4129–4133. Reykjavik: European Language Resources Association (ELRA).