Interpreter identification in the Polish Interpreting Corpus

Danijel Koržinek
Agnieszka Chmiel

Danijel Koržinek
Polish-Japanese Academy of Information Technology,
Warsaw, Poland
danijel@pja.edu.pl;
ORCID:
0000-0002-2916-4856

Agnieszka Chmiel
Adam Mickiewicz University,
Poznań, Poland
achmiel@amu.edu.pl;
ORCID:
0000-0002-2138-3974

Abstract

This paper describes automated identification of interpreter voices in the Polish Interpreting Corpus (PINC). After collecting a set of voice samples of interpreters, a deep neural network model was used to match all the utterances from the corpus with specific individuals. The final result is very accurate and provides a considerable saving of time and accuracy off human judgment.

Keywords: automatic speaker identification; speech corpora; speaker annotation; interpreting corpora; European Parliament

Resum

Aquest article descriu la identificació automatitzada de veus d'intèrprets al Corpus d'Intèrprets Polonès (Polish Interpreting Corpus, PINC). Després de recollir un conjunt de mostres de veu de diversos intèrprets, s’ha utilitzat un model de xarxa neuronal profunda per fer coincidir les mostres de parla del corpus amb les de cada individu. El resultat final és molt precís i proporciona un estalvi considerable de temps i de precisió en la interpretació humana.

Paraules clau: identificació automàtica de veus; corpus de parla; transcripció de veus; corpus d’intèrprets; Parlament Europeu

Resumen

Este artículo describe la identificación automática de voces de intérpretes en el Corpus Polaco de Interpretación. Tras recopilar una serie de muestras de voces de intérpretes, se utilizó un modelo de red neuronal profunda para asociar todas las elocuciones del corpus con individuos específicos. El resultado final es muy acertado, lo cual implica un ahorro considerable de tiempo y análisis humano.

Palabras clave: identificación automática de voces; corpus de discursos; anotación de hablantes; corpus de interpretación; Parlamento Europeo

Rebuda: 4 de febrer de 2021 / Acceptació: 26 d’agost de 2021 / Publicació avançada: 15 de desembre de 2021

Revista Tradumàtica. Número 19 (desembre 2021).
Cita recomanada: KORŽINEK, DANIJEL; CHMIEL, AGNIESZKA; (2021). «Interpreter identification in the Polish Interpreting Corpus». Revista Tradumàtica. Tecnologies de la Traducció, 19, 276-288.
<https://doi.org/10.5565/rev/tradumatica.291>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

1. Introduction

Interpreting corpora offer excellent research opportunities thanks to numerous automatically generated metadata that can provide insights into various aspects of interpreters' output. These include for instance information about speaking speed, parts of speech, numbers of word types and tokens. Nevertheless, there are many aspects of speech records that are not easily processed automatically and require laborious human processing. As a result, many scholars refrain from obtaining such metadata and do not include them in their analyses. One such aspect is interpreter voice identification since interpreter identity information (unlike that of the speaker) is not always available for materials used to create interpreting corpora. Metadata regarding interpreter voice could be useful to study individual differences in processing and interpreting styles. It can also be helpful in controlling for individual variation in statistical modelling. To fill in this niche, this article presents the procedure of interpreter voice identification in PINC (Polish Interpreting Corpus) that included both human and automatic processing.

There are several levels of automated speaker annotation available:

1. *Speaker change detection* is used to detect the point in time where one speaker is replaced by another, without analysing anything about their identities.
2. *Speaker diarization* is used to annotate all the segments of individual speakers without inferring their identity (that is using random or sequential labels).
3. *Speaker recognition or identification* is used to annotate segments of audio with actual identities (that is name and surname) of individual speakers. This process requires a set of labelled recordings used to train the models used for identifying speakers.

This article deals with the third problem.

The procedure of interpreter voice identification was performed on the Polish Interpreting Corpus (PINC), which is a corpus of original Polish or English speeches from the European Parliament and their respective interpretations into English or Polish. Alongside other corpora that make use of speeches and interpretations available from the European Parliament website, such as TIC (Kajzer-Wietrzny, 2012), EPICG (Defrancq, Plevoets and Magnifico, 2015) and EPTIC (Ferraresi and Bernardini, 2019), it belongs to the EPIC suite of corpora (Bernardini et al., 2018). PINC includes 5 subcorpora given in the table below.

	Number of speeches	Number of tokens
English source texts	230	54,090
Interpretations into Polish	230	36,649
Polish source texts	290	51,078
Interpretations into English (retour – into B language)	232	41,283
Interpretations into English (into A language)	58	10,578

Table 1. Structure of the Polish Interpreting Corpus (PINC).

The subcorpora include English source texts and their interpretations into Polish as well as Polish source texts and their interpretations into English, performed either by Polish interpreters providing retour interpretations (i.e., into their foreign or B language), or by English interpreters working from their passive or C language into their native or A language. In its final form, PINC includes over 190,000 tokens. Thanks to the presence of the retour subcorpus, it is possible to compare performances of the same interpreters working in opposing directions: from B to A (English-Polish) and from A to B (Polish-English).

The European Parliament website provides audio-visual content and clearly identifies the original speakers – MEPs or other individuals participating in the proceedings (such as commissioners or guests). However, there is no information on the identity of interpreters, and this has to be established solely on the basis of auditory data. Interpreting corpora that feature interpreter voice identification typically include target texts produced by few interpreters. For instance, the CEIPPC (Corpus of Chinese-English Interpreting for Premier Press Conferences) built by Wang (2012) includes 7 interpreters. Similarly, the EUDEB14 corpus of EU presidential debates broadcast in Italy includes contributions from 7 interpreters (Dal Fovo, 2018). EPTIC creators acknowledge the fact that interpretations into English in the European Parliament are either provided by the English booth (i.e., by native speakers of English into their A language) or by non-native speakers, and EPTIC metadata include only two pieces of information about the interpreter: gender and whether the interpretation is delivered into the native or foreign language (https://corpora.dipintra.it/eptic/?section=documentation). Additionally, regional variety of the language used by the interpreter may be determined, as is the case in EPICG, given the binational nature of the Dutch booth (Bernardini et al., 2018).

Interpreter voice identification has been previously used by scholars who have explored gender differences in interpreters' output (Russo, 2016, Magnifico and Defrancq, 2016, Collard and Defrancq, 2020, Russo, 2018). However, to the best of our knowledge, not many studies to date have included individual information about interpreters in their analyses to either show or monitor individual differences. A noticeable example would be the NAIST Simultaneous Translation Corpus (Neubig et al., 2018) in which the interpreters' experience was analysed. However, this corpus did not include fully naturalistic data as three interpreters were hired to interpret a previously collected corpus of source texts. We can thus assume that the interpretations were created in an experimental setting and do not originate from real-life interpreter-mediated events.

Detailed identification of interpreter voices in a corpus presents interesting research and analytical opportunities. In our case, it is important to recognise the interpreters' identities because we would like to correlate PINC data with other data from the same interpreters obtained in experimental studies (Chmiel, 2012, 2016, 2018). The studies by Chmiel referred to above include data on working memory spans, word translation latencies and reaction times in a priming task – all are indexes of cognitive or lexical processing in interpreting and may be used in advanced correlations with the PINC data (for instance on fluency, pauses, compression rates, accuracy, etc.).

Interpreter identification data may also be used to examine individual differences in various aspects of processing in interpreting. For instance, we have already used such data to show individual differences in text compression rates in Polish-English retour interpretations (Chmiel et al., submitted). With precise interpreter identification it is possible to adopt a within-group study design. In the case of PINC, we may compare interpreter performance when interpreting into A language and when providing retour interpretation (into B language). A within-group design (i.e., data from the same group of interpreters for both interpreting directions) is advantageous as it has greater statistical power than a between-groups design (i.e., data from two different groups of interpreters). Finally, interpreter identity may be used as a random factor in statistical models to better capture data variance.

2. Speaker identification in recorded speech – state of the art

The task of identifying speakers in recorded speech has a long history and can have many practical applications. One that immediately comes to mind is security and access control (Bergl et al., 2001), where the problem is usually framed as speaker verification, that is, confirming the identity of a specific individual based on their voice sample. This can be done either in a text dependent manner (Zhang et al. 2019), requiring the person to speak a specific sequence of words, or a text independent manner (Torfi et al., 2018), which works regardless of what utterance was spoken by the speaker. In either case, the problem can be regarded as a form of biometric analysis, where the onus is on the quality of the audio (pitch, timbre, speaking style) rather than the content. This is significantly different from tasks like speech recognition, where the goal is to infer the content regardless of who the speaker is. This highlights how different the speaker recognition tools and processing pipelines are to those involved in other, more common speech analysis tasks. The American National Institute for Standards and Technology (NIST) has organised the Speaker Recognition Evaluation competition since 1996 and in 2018 they managed to attract 48 different teams from around the globe (Sadjadi et al., 2019).

Another common use case involves analysing a speech recording that contains two or more speakers taking turns in a conversation. The purpose of performing speaker identification in such cases may be to improve the performance of other systems, for example by selecting appropriate speaker-adapted speech recognition models. In some cases, it may go even as far as performing an audio speech separation of both speakers in a single mono recording (Pariente et al., 2020). It is worth mentioning that the identity of the individuals is not always available or necessary, so there is another class of problem known as speaker diarization which attempts to recognize individuals in an unsupervised manner (Sell et al., 2018), that is without requiring any labelled data to train the system on. Diarization is used to annotate the different speakers using random labels (like speaker #1, speaker #2, etc) and it does so completely automatically, without having any prior information about their voices. Speaker identification, on the other hand, requires a labelled set of recordings for each voice we want to detect.

The purpose for using speaker identification in this article was to add a layer of annotation to an already existing dataset. Historically, some of the earliest approaches used algebraic mathematical models to describe the database of individuals as a vector space. Inspired by the now famous Eigenfaces paper (Turk, Pentland, 1991), which described a method for the identification of people by their portraits, the Eigenvoice method (Kuhn et al., 1998) used principal component analysis to determine the optimal vector representation of speakers. Over time this culminated in the method known as the i-vector (Dehak et al., 2010) which stood as the most popular approach until the advent of deep learning. One of the attractive features of i-vectors was the ability to describe any audio sample as a multidimensional real vector. Such an approach is currently known as an embedding and is used in many other fields like natural language processing (e.g., word2vec, sent2vec). This idea was later kept in the deep-learning based methods, like d-vector (Variani et al., 2014) and X-vector (Snyder et al., 2018).

3. Interpreter voice identification in PINC – tools and procedures

In this article, the analysis is based on the X-vector method for speaker identification. This method relies on a multilayer deep neural network trained in a discriminative manner. The network we used consisted of five layers of Time-Delay Neural Network (TDNN) topology, followed by a statistical pooling layer and then followed by three dense layers. The last layer was trained to classify a large database of speakers, but the final network used the embedding derived from the third to the last layer – a trick often used in transfer learning. This embedding contains 512 real values that can be calculated for any audio provided to the network.

The speaker identification algorithm did not end there. The output of the network was further processed to improve the final performance of the identification process. First, a global mean computed on a large sample of speakers was subtracted from the data during training and inference. Next, the number of dimensions was further reduced to 128 using an LDA transform. Following this, the vectors were normalized to a unit norm (i.e., mapped to a hyper-sphere of diameter 1) as this reduces the variability in data coming from changes in energy and different SNR levels. This is all standard practice for both i-vectors and X-vectors (García-Romero and Espy-Wilson, 2011). Finally, the PLDA algorithm was used to perform actual classification. Before the PLDA algorithm was applied, one can imagine that each voice sample was converted into a 128 element vector. The PLDA algorithm can both classify the sample to a voice in the database and provide a score that denotes the quality of such a match. This allowed us to reject some samples as not matching at all if the score value is too low.

The model used here was pretrained on publicly available datasets called VoxCeleb 1 and VoxCeleb 2 (Nagrani et al., 2017). The data was further augmented to anticipate various acoustic environments. The whole process is outlined in Snyder et al. (2018). To our knowledge, these datasets do not contain any parliamentary data, yet the model was sufficiently generic to perform well with our data.

The data used in the analysis is divided into two subsets: the enrolment set and the test set. The former included manually annotated files to identify individual interpreters. The enrolment set was prepared independently using known samples of peoples' voices based mostly on interpreter voice samples collected by Kajzer-Wietrzny in the TIC corpus (2012). The files were also verified by co-workers with personal connections with those individuals. It contained 57 voice samples for 32 different speakers, each speaker having between 1 and 4 samples. Each sample was between 21 and 192 seconds, averaging at 87 with a standard deviation of 35.7.

The test set was the actual PINC corpus set of audio files that required interpreter voice identification. It was known to contain mostly the same interpreters as those in the enrolment set, but some exceptions were possible, meaning that certain files would contain voices not present in the enrolment database. The test set contained 462 files of various lengths, truncated to 30 seconds, as described below. Initially, the human compilers who selected speeches for PINC used the enrolment set and their subjective judgment on audio files to match the interpretations with the samples. If the interpreter voice was judged as not matching any samples, the enrolment set was expanded to include a new sample. Problematic voices were consulted with all compilers but not all voices could be matched to the samples with sufficient confidence. We thus decided to rely on automatic interpreter voice identification.

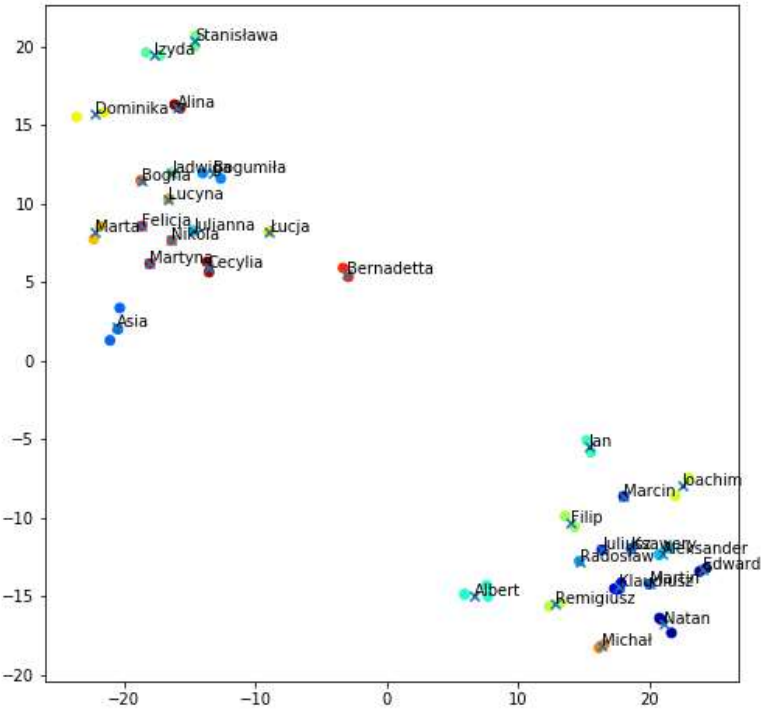

Figure 1. A t-SNE representation of X-vectors present in the enrolment subset. Each dot represents one voice sample, and the colour represents the identity. A mean vector was calculated for every individual, marked with a cross and labelled using their pseudo-identity.

To observe and control the whole process, we used a visualization tool for presenting the speaker vectors in a 2D scatter plot using the t-SNE algorithm (Van der Maaten et al., 2008). This obviously led to a drastic compression of the number of dimensions, but it is still a popular approach for visualization. Figure 1 shows each voice sample as a single dot in the graph. The names on the graph were randomly assigned. They do not denote the speakers' actual identities. Next, we computed the mean of the vectors belonging to a particular individual, thus each voice was described by a single vector and stored in our database.

In the following step, we took all the files from the PINC database and extracted the vectors representing the speakers in those files. There were several steps that needed to be taken first, however. The files available on the European Parliament website were not perfectly clear and contained some audio interference at the start and end of the recording (other interpreters or speakers in the Parliament). In order to streamline the process, it was decided to take 30 seconds of the roughly middle portion of each recording instead of the full audio. Next, an essential step was to perform Voice Activity Detection on the audio to reject any portions that did not contain actual speech. The X-vector extractor would generate random noise in such cases and negatively affect the final outcome. Since the audio from interpreters came from a soundproof booth, there was very little (if any) background noise, so we opted for a simple energy based VAD (Voice Activity Detection) solution. Others, who intend to use this approach on more complicated data should use a more elaborate, possibly DNN-based VAD solution. After extracting the X-vectors from these files, we plotted them again as one dot per file using the same t-SNE embedding as before.

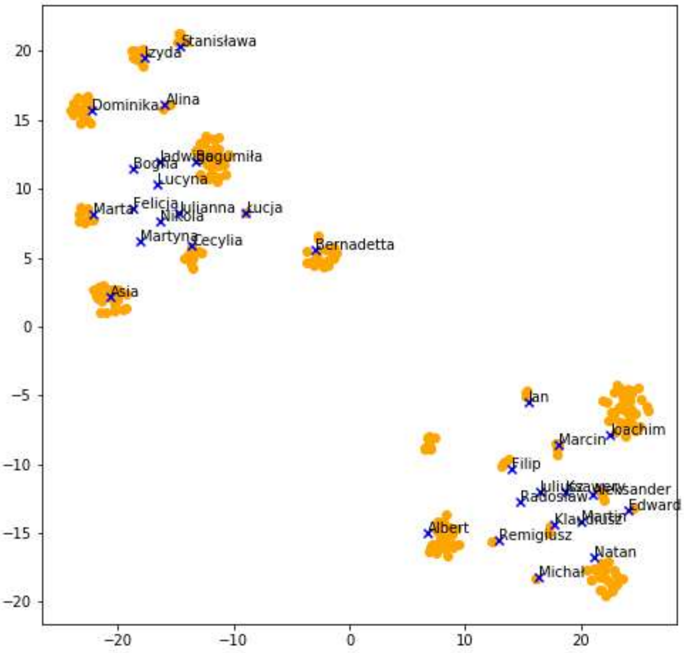

Figure 2. A t-SNE representation of the X-vectors of all the recordings in the test subset, i.e., the PINC database. The mean vectors and their names from Figure 1 are overlaid on top to show the proximity between the speakers and the files undergoing classification.

Figure 2 shows all the initially unidentified files using the same colour. The same vector locations from the previous plot were overlaid on top of those files. The last step of our identification process was to match the files to the speakers. This was done by computing a PLDA score for each speaker-file pair and for each file we chose the speaker whose score was the highest. The final result was prepared in the form of a table, similar to the one in Table 1. This allowed for a more informed process of verification. The human judge could then take both the score of the best speaker as well as other potential candidates into account when correcting the final annotation.

Table 2 below shows a small sample of the classification results matching each test utterance to each speaker in the database. The best speaker and corresponding score are given in the second and third columns, followed by a score for all other speakers, in case we need to dispute the best choice provided by the system. The colour is used to describe the different value ranges (from dark green being the lowest, to dark orange being the highest values) to quickly identify maxima in each row or column.

Utterance	Best speaker	Best score	*Alina*	*Bogumiła*	*Bernadetta*	*Cecylia*	*Albert*	*Edward*
PL0001_en	Marta	62.48	18.19	5.13	-29.60	11.45	-88.16	-2.53
PL0002_en	Marta	71.47	18.58	3.66	-30.03	11.22	-92.95	-12.71
PL0003_en	Michał	48.19	-103.88	-112.88	-82.80	-76.02	-12.35	-69.14
PL0004_en	Michał	44.68	-78.19	-89.67	-64.00	-69.47	-19.05	-68.93
PL0005_en	Remigiusz	46.26	-98.34	-92.65	-72.74	-64.72	-4.59	-56.02
PL0006_en	Remigiusz	45.55	-96.34	-93.48	-66.59	-58.85	-2.46	-58.71
PL0007_en	Martyna	65.14	9.78	-6.80	-29.89	12.05	-90.71	-21.05
PL0008_en	Jan	41.07	-99.06	-72.36	-55.11	-57.01	-15.67	-39.40
PL0009_en	Jan	55.61	-105.01	-84.37	-61.51	-62.29	-1.94	-41.13

Table 2. Sample classification results

After obtaining the automatically generated interpreter identification data, we compared it against the initial subjective judgments of human annotators. We found a 15% mismatch. Each mismatched case was then revisited by the same annotators. Automatic results were adopted as final. Given the novelty of our dataset, it is difficult to compare our findings directly with those in other publications, which usually strive to solve more challenging examples. That is why we decided to compare the method described in this paper with the i-vector approach, which is very well known and often used as a baseline in similar situations. The final result of the comparison is provided in Table 3. The dataset had a total of 13 utterances which contained voices that were not in the enrolment set. These were counted to the classification error rate. The AUC (area under the curve) and EER (equal error rate) scores were counted for recognized classes only, so the unknown individuals were omitted in that statistic. Those metrics are both derived from the ROC (receiver operator characteristic) curve which is a graph representing the performance of the system as a trade-off between its type I and type II errors. The AUC is the area beneath the ROC curve, while EER is the value on the graph where both type I and type II errors are equal. The results are consistent with other literature (Snyder et al., 2018).

Method	Error count	Error rate	EER	AUC
i-vector	18	7.83%	2.24%	99.89%
X-vector	14	6.87%	1.39%	99.96%

Table 3. Results of the performance of two speaker identification methods used on our parliamentary corpus when compared to human experts. Expressed in 3 common measures: classification error rate, equal error rate (EER) and area under the curve (AUC).

5. Technical details

The experiments in this article were performed using the Kaldi toolkit (Povey et al., 2011). The project is available on the official Github repository (https://github.com/kaldi-asr/kaldi). Alternatively, one can also utilise a pre-compiled image using the Docker environment hosted on the public DockerHub repository (https://hub.docker.com/r/kaldiasr/kaldi). The models used for this study are available on the project's official model page (http://kaldi-asr.org/models.html). We used the model with the code M8 which seemed to be trained on the largest amount of data (at the time of publishing). From there we used the model version 1a, which is the XVector model, as well as the i-vector model, which was used only for comparison in the final results.

The whole experiment procedure consisted of the following steps:

1. Installation of the toolkit
2. Creation of a project directory, similar to the *egs/sitw/v2* example available in the toolkit
3. Creation of the data folders containing the description of all the audio files being analysed – separately for the enrolment and test subsets
4. Feature extraction
5. Voice activity detection
6. X-vector extraction
7. PLDA scoring

A detailed step-by-step tutorial is provided on our Github project page (https://github.com/PINC-Project).

6. Conclusions

Several conclusions can be drawn from the automatic classification exercise described above. Initially, the match seemed to be pretty robust: there were individuals that were fairly close to each other in the speaker space, but there did not seem to be any utterances that could not be matched due to the proximity of the individuals. There were several individuals that did not have any utterances associated with them as well as utterances that did not seem to have any speakers in close proximity. The latter could be detected by having a negative PLDA score.

We hope that this paper, and especially the step-by-step tutorial it refers to, will be useful to other scholars working on processing interpreting and other speech corpora. As pinpointed above, metadata including interpreter voice identity can be useful in analyses in various corpus-based studies and makes it possible to adopt a within-group, i.e., a more powerful study design. Our results showed a small mismatch between human and automatic processing. All the mismatched cases were resolved in favour of automatic processing, which shows how successful such a procedure can be. Undoubtedly, the main advantage of using the automatic speaker identification was the time saved in preparing the annotation. The experts that performed the initial manual annotation noted how difficult and time-consuming it was to match hundreds of recordings to several dozens of speakers. We do not have detailed data on the exact time spent by human annotators trying to determine which recordings belonged to the same interpreter, nevertheless, it was an effortful task as it required repeated listening to the annotated recording and previously annotated recordings to directly compare the voices and determine if they were similar. Conversely, the task of verifying the automatic results was much quicker and easier.

Acknowledgments

This research was supported by the UMO-2018/30/E/HS2/00035 (SONATA BIS 8) research grant funded by the National Science Centre (2019-2024). Some of the tools used were developed as part of the Clarin-PL project.

References

Bergl, Vladimir; *et al.* (2001). *Apparatus and methods for user identification to deny access or service to unauthorized users*. U.S. Patent No. 6246751. 12 Jun. 2001. <https://patents.justia.com/patent/6246751>. [Accessed: 20211116].

Bernardini, S.; Ferraresi, A.; Russo, M.; Collard, C.; Defrancq, B. (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In: Russo, M.; Bendazzoli, C.; Defrancq, B. (eds.). *Making way in corpus-based interpreting studies.* Singapore: Springer singapore, pp. 21-42. <https://doi.org/10.1007/978-981-10-6199-8_2>. [Accessed: 20211116].

Chmiel, A. (2012). Pamięć operacyjna tłumaczy konferencyjnych mierzona metodą RSPAN. In: Piotrowska, M. (ed.). *Kompetencje tłumacza*. Kraków: Tertium, pp. 137-154.

Chmiel, A. (2016). Directionality and context effects in word translation tasks performed by conference interpreters. *Poznan Studies in Contemporary Linguistics,* v. 52, n. 2, pp. 269–295. <https://doi.org/10.1515/psicl-2016-0010>. [Accessed: 20211116].

Chmiel, A. (2018). Meaning and words in the conference interpreter's mind: Effects of interpreter training and experience in a semantic priming study. *Translation, Cognition & Behavior,* v. 1, n. 1, pp. 21–41. <https://doi.org/10.1075/tcb.00002.chm>. [Accessed: 20211116].

Chmiel, A.; Kajzer-Wietrzny, M.; Koržinek, D.; Janikowski, P.; Jakubowski, D.; Polakowska, D. (2019). Fluency parameters in the Polish Interpreting Corpus (PINC). In: Kajzer-Wietrzny, M.; Bernardini, S.; Ferraresi, A.; Ivaska, I. (eds.). *Empirical investigations into the forms of mediated discourse at the European Parliament: A thematic session at the 49th Poznań Linguistic Meeting (PLM2019).* <http://wa.amu.edu.pl/~wjarek/PLM2019/PLM2019_Thematic_session_Mediated_discourse_European_Parliament.pdf>. [Accessed: 20211116].

Collard, C.; Defrancq, B. (2020). Disfluencies in simultaneous interpreting: A corpus-based study with special reference to sex. In: Defrancq, B.; Vandevoorde, L.; Daems, J. (eds.). *New empirical perspectives on translation and interpreting.* London: Routledge, pp. 264-299. <https://doi.org/10.4324/9780429030376-12>. [Accessed: 20211116].

Dal Fovo, E. (2018). European Union Politics Interpreted on Screen: A corpus-based investigation on the interpretation of the third 2014 EU presidential debate. In: Russo, M.; Bendazzoli, C.; Defrancq, B. (eds.). *Making way in corpus-based interpreting studies*. Singapore: Springer Singapore, pp. 157-184. <https://doi.org/10.1007/978-981-10-6199-8_9>. [Accessed: 20211116].

Defrancq, B.; Plevoets, K.; Magnifico, C. (2015). Connective Items in Interpreting and Translation: Where Do They Come From?. In: Romero-Trillo, J. (ed.). *Yearbook of Corpus Linguistics and Pragmatics 2015*. Cham: Springer, pp. 195–222. <https://doi.org/10.1007/978-3-319-17948-3_9>. [Accessed: 20211116].

Dehak, N.; *et al.* (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing,* v. 19, n. 4, pp. 788-798. <https://doi.org/10.1109/TASL.2010.2064307>. [Accessed: 20211116].

Ferraresi, A.; Bernardini, S. (2019). Building EPTIC. In: Doval, I.; Sánchez Nieto, M.T. (eds.). *Parallel Corpora for Contrastive and Translation Studies: New resources and applications.* Amsterdam: John Benjamins. (Studies in Corpus Linguistics; 90), pp. 123-139. <https://doi.org/10.1075/scl.90.08fer>. [Accessed: 20211116].

Garcia-Romero, D.; Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. Conference in: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31. In: *DBLP Computer Science Bibliography.* <https://dblp.uni-trier.de/db/conf/interspeech/interspeech2011.html#Garcia-RomeroE11>. [Accessed: 20211116].

Kajzer-Wietrzny, M. (2012). Interpreting universals and interpreting style [PhD. Thesis]. Uniwersytet im. Adama Mickiewicza w Poznaniu, Poznán. Unpublished.

Kuhn, R.; *et al.* (1998). Eigenfaces and eigenvoices: Dimensionality reduction for specialized pattern recognition. In: *1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No. 98EX175).* <https://doi.org/10.1109/MMSP.1998.738915>. [Accessed: 20211116].

Magnifico, C.; Defrancq, B. (2016). Impoliteness in interpreting: A question of gender? *Translation and Interpreting,* v. 8, n. 2, pp. 26-45. <http://www.trans-int.org/index.php/transint/issue/view/40>. [Accessed: 20211116].

Nagrani, A.; Chung, J.S.; Zisserman, A. (2017). VoxCeleb: A Large-Scale Speaker Identification Dataset. In: *Proc. Interspeech 2017*, pp. 2616-2620. <https://doi.org/10.21437/Interspeech.2017-950>. [Accessed: 20211117].

Neubig, G.; Shimizu, H.; Sakti, S.; Nakamura, S.; Toda, T. (2018). The NAIST Simultaneous Translation Corpus. In: Russo, M.; Bendazzoli, C.; Defrancq, B. (eds.). *Making Way in Corpus-based Interpreting Studies.* Singapore: Springer Singapore, pp. 205-215. <https://doi.org/10.1007/978-981-10-6199-8_11>. [Accessed: 20211117].

Pariente, M.; Cornell, S.; Deleforge, A.; Vincent, E. (2020). Filterbank design for end-to-end speech separation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. <https://doi.org/10.1109/ICASSP40776.2020.9053038>. [Accessed: 20211117].

Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; Silovsky, J.; Stemmer, G.; Vesely, K. (2011). The Kaldi speech recognition toolkit. In: *IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF).* IEEE. <https://www.danielpovey.com/files/2011_asru_kaldi.pdf>. [Accessed: 20211117].

Russo, M. (2016). Orality and Gender: A corpus-based study on lexical patterns in simultaneous interpreting. *MonTI, Monografías de Traducción e Interpretación*, Special Issue 3, pp. 307-322. <https://doi.org/10.6035/MonTI.2016.ne3.11>. [Accessed: 20211117].

Russo, M. (2018). Speaking Patterns and Gender in the European Parliament Interpreting Corpus: A Quantitative Study as a Premise for Qualitative Investigations. In: Russo, M.; Bendazzoli, C.; Defrancq, B. (eds.). *Making Way in Corpus-based Interpreting Studies.* Singapore: Springer Singapore. (New Frontiers in Translation Studies), pp. 115-131. <https://link.springer.com/book/10.1007/978-981-10-6199-8>. [Accessed: 20211117].

Sadjadi, S.O.; Greenberg, C.; Singer, E.; Reynolds, D.; Mason, L.; Hernandez-Cordero, J. (2019). *The 2018 NIST Speaker Recognition Evaluation*. In: *Proc. Interspeech 2019, pp.* 1483-1487. <https://doi.org/10.21437/Interspeech.2019-1351>. [Accessed: 20211117].

Sell, G.; Snyder, D.; McCree, A.; Garcia-Romero, D.; Villalba, J.; Maciejewski, M.; Manohar, V.; Dehak, N.; Povey, D.; Watanabe, S.; Khudanpur, S. (2018). Diarization is Hard:

Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In: *Proc. Interspeech 2018,* p. 2808-2812. <https://doi.org/10.21437/Interspeech.2018-1893>. [Accessed: 20211117].

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In: 2*018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 5329-5333. <https://doi.org/10.1109/ICASSP.2018.8461375>. [Accessed: 20211117].

Torfi, A.; Dawson, J.; Nasrabadi, N. M. (2018). Text-independent speaker verification using 3d convolutional neural networks. In: *2018 IEEE International Conference on Multimedia and Expo (ICME).* IEEE, pp. 1-6. <https://doi.org/10.1109/ICME.2018.8486441>. [Accessed: 20211117].

Turk, M. A.; Pentland, A. P. (1991). Face recognition using eigenfaces. In: Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition IEEE, pp. 586-587. <https://doi.org/10.1109/CVPR.1991.139758>. [Accessed: 20211117].

Van der Maaten, L.; Hinton, G. (2008). Visualizing data using t-SNE. J*ournal of Machine Learning Research,* v. 9, pp. 2579-2605. <https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>. [Accessed: 20211117].

Variani, E.; Lei, X.; McDermott, E.; Moreno, I. L.; Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 4052-4056. <https://doi.org/10.1109/ICASSP.2014.6854363>. [Acccessed: 20211117].

Wang, B. (2012). A descriptive study of norms in interpreting: Based on the Chinese-English consecutive interpreting corpus of Chinese premier press conferences. *Meta: journal des traducteurs = Meta: Translators' Journal*, v. 57, n. 1, pp. 198-212. <https://doi.org/10.7202/1012749ar>. [Accessed: 20211117].

Zhang, Y.; Yu, M.; Li, N.; Yu, C.; Cui, J.; Yu, D. (2019). Seq2seq attentional siamese neural networks for text-dependent speaker verification. In: I*CASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 6131-6135. <https://doi.org/10.1109/ICASSP.2019.8682676>. [Accessed: 20211117].