Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1193–1200
Marseille, 20-25 June 2022

EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus

Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, Elke Teich
Saarland University
Campus A2.2, 66123 Saarbrücken, Germany
{heike.przybyl, stefan.fischer}@uni-saarland.de
{e.lapshinova, k.menzel, e.teich}@mx.uni-saarland.de

Abstract

In this paper, we describe the creation and annotation of EPIC UdS, a multilingual corpus of simultaneous interpreting for English, German and Spanish. We give an overview of the comparable and parallel, aligned corpus variants and explore various applications of the corpus. What makes EPIC UdS relevant is that it is one of the rare interpreting corpora that includes transcripts suitable for research on more than one language pair and on interpreting with regard to German. It not only contains transcribed speeches, but also rich metadata and fine-grained linguistic annotations tailored for diverse applications across a broad range of linguistic subfields.

Keywords: Simultaneous interpreting corpus, European Parliament data, multilingual language resource, corpus annotation

1. Introduction

The European Parliament (EP) has been described as "one of the most promising sources of interpreting and intermodal corpora" for European languages due to the availability, accessibility and controlability of data and the professional interpreter status (Bernardini et al., 2018). Various interpreting corpus projects make use of this rich data source for interpreted speech. At the same time, the EP is a rather specific interpreting setting. EU interpreters work against the backdrop of extreme multilingualism in large meetings (Graves et al., 2021) and the EP context is not necessarily entirely representative for all interpreting services provided in the EU.

Due to the generally time-consuming nature of spoken language corpora compilation, simultaneous interpreting corpora are still rare. They exist for a restricted set of language combinations, and most of them can be classified as nano-corpora (Defrancq, 2018). New interpreting corpora are currently being compiled with different goals and users in mind, and thus with varying transcription levels, meta-data attributes and with different degrees of availability. These corpus projects are highly relevant as interpretese still remains a largely understudies phenomenon. We briefly discuss a non-exhaustive list of examples in this paper (Section 2).

Using the standards established for existing resources, the EPIC UdS (European Parliament Interpreting Corpus at Saarland University) project was designed in a way that its transcriptions can be used in combination with data available within previously compiled EPIC corpora for other language combinations (cf. Section 3.2). Furthermore, we provide various corpus variants that can be used for different research foci, e.g. one corpus variant including disfluencies such as filled and silent pauses for studying cognitive aspects of interpreting vs. a "cleaned" corpus variant without disfluencies enabling more accurate linguistic annotations needed for research on specific lexical and syntactic features.

The remainder of the paper is organized as follows: In Section 2, we provide an overview of other existing interpreting corpora and discuss related work. Section 3 describes procedures of data collection, transcription and annotation. In Section 4, we present the existing corpus variants. Possible applications are discussed in Section 5. We conclude with a summary and outlook in Section 6.

2. Related Work

2.1. Existing Corpora

The EPIC UdS corpus can be situated in the European Parliament Interpreting family of corpora which was initiated with EPIC, created at Bologna university for Italian, English and Spanish (Russo et al., 2005). Other EP corpora include EPICG, covering English, French, Dutch and Spanish (Defrancq et al., 2015), PINC for Polish and English (Chmiel et al., 2021) as well as EP interpreting corpora efforts in Belgrade, Louvain and Lisbon (Bernardini et al., 2018). Recently, the corpus ESIC has been released with EP speeches in English and simultaneous interpreting into Czech and German (Macháček et al., 2021). The various European Parliament interpreting corpora mainly differ with regard to the languages covered, selection of interpreter status (interpreting into assumed native language e.g. EPIC, EPICG and EPIC UdS, or including "return" for PINC) and transcription post-processing such as time-alignment for EPICG with EXMARaLDA or sentence alignment for EPIC as well as the sentence aligned and dependency parsed variants for EPIC UdS. EPIC UdS adds German (as an additional widely used language spoken in the EP) to this corpus family, referred to as "EPIC suite of corpora" by (Bernardini et al., 2018).

Availability issues of interpreting data lead to the fact that most existing interpreting resources are of political nature. Beside the aforementioned EPIC family of corpora, SIREN (Dayter, 2018), an English-Russian corpus of speeches and interpretations held at the United Nations and CEPIC (Pan, 2019), a large-scale (6.5 mio tokens) Chinese-English corpus, can also be placed in the political context.

Interpreting corpora that include texts from other genres and domains include the DIRSI-C, a parallel, aligned corpus of English-Italian transcripts of international conferences, following the same methodology as the EPIC family (Bendazzoli, 2010), the WAW corpus in Arabic and English also with texts from international conferences (Temnikova et al., 2017) and HeiCIC, the Heidelberg Conference Interpreting Corpus (Kunz et al., 2021). HeiCIC contains authentic speeches by scientists and experts presenting their research on a variety of topics including electrical engineering in car manufacturing, astronomy, investor relations and annual general meetings of international corporations. These speeches were simultaneously interpreted by learners and professionals in eight languages. The English-German core corpus contains several parallel interpretations by students and professionals with different levels of interpreter expertise. Apart from EPIC UdS and ESIC, HeiCIC is the only other interpreting corpus we are aware of that includes German in its language combinations.

Further corpora that include texts from other genres and domains are NAIST (Doi et al., 2021) and BST (Zhang et al., 2021). The former is a English-Japanese interpreting corpus containing simultaneous interpreting of academic lectures, speeches on everyday topics, press conferences by politicians to business representatives and talks from TED conferences. The latter covers the English-Chinese language pair. It contains speeches and their interpretations involving a wide range of domains, including IT, economy, culture, biology, arts and others.

2.2. Studies of Interpreting

Corpus-based studies of interpreting often focus on the analysis of interpretese, i.e. specific features of interpreted texts that distinguish them not only from spoken originals but also from written translation ((Sandrelli and Bendazzoli, 2005; Bernardini et al., 2016; Ferraresi and Miličević, 2017) and (Defrancq et al., 2015; Kajzer-Wietrzny, 2012; Dayter, 2018)). Such studies use linguistic patterns to quantitatively and qualitatively analyse the interpretese phenomena in the language corpora at hand.

Interpreting corpora are also essential for computational studies of interpretese which use automatically extracted features to automatically tease apart interpreting from other language products. For instance, He et al. (2016) use both shallow surface features and linguistically motivated features such as passive constructions, general nouns, etc. to automatically distinguish between interpreted and translated texts. Bizzoni and Teich (2019) use interpreting and translation corpora to create bilingual word embedding spaces that they explore for differences. Lapshinova-Koltunski (2021) uses a number of hand-crafted features inspired by variational linguistics to automatically classify interpretations and translations, as well as comparable spoken and written non-translations. In a recent study, Lapshinova-Koltunski et al. (2021a) use comparable subcorpora of EPIC and Europarl UdS in a classification task to show that interpreting is the most distinctive type of language production. Moreover, they analysed the translationese and interpretese features in the data and found both general translationese effects and effects that are unique to interpreting and translation, respectively.

And last but not least, interpreting corpora are used as training and test data for automatic speech translation, where the tasks include simultaneous and offline speech translation, multilingual as well as low-resourced speech translation, see proceedings of the series of the International Conference on Spoken Language Translation (IWSLT), such as (Federico et al., 2021; Federico et al., 2020) and previous editions.

3. Corpus Compilation

3.1. Data Collection

As manual transcription of spoken data is very time-consuming, we used some existing European Parliament transcripts as a starting point for our corpus data: The main part of the English subcorpora – English spoken originals (EN ORG) and English interpretations with German and Spanish as source (SI DE EN and SI ES EN) – are taken from TIC (Kajzer-Wietrzny, 2012) and individual English original speeches from EPICG (Defrancq et al., 2015). In order to ensure comparability of data, these existing transcripts were revised, and the newly transcribed dataset for German and Spanish, compiled at Saarland University (UdS), were transcribed according to transcription guidelines based on EPICG (Bernardini et al., 2018).

EPIC UdS covers speeches held by MEPs (members of the European Parliament) between 2008 and 2013. Original speakers were selected based on their nationality. Native language was operationalised via speaker's nationality matching the language of the original speech: e.g., if German or Austrian MEPs delivered their speech in German, it was assumed that they are speaking in their native language (Kajzer-Wietrzny, 2015). As for interpreters' native language, the language combinations that are covered in this corpus (DE <> EN and ES > EN) are generally interpreted from a language which the interpreter understands perfectly or in which the interpreter is perfectly fluent (but into which they do not work) to the interpreter's mother tongue or its strict equivalent. Additionally, for these widely used languages in the EP, usually no relay interpreting via a third "pivot" language is used. A fairly reliable predictor of relay interpreting is Ear-Voice-Span (EVS, the lag time between original speakers utterance and interpreter production) being constantly above four seconds (Bernardini et al., 2018). EPIC UdS is not time-aligned. However, the multilingual videos used for transcription allow for manual validation of EVS not being above four seconds over a longer period of time. We therefore assume that interpreters speak in their native language, interpreting directly from the source language.

3.2. Transcription

Automatic speech recognition and transcription to text has advanced hugely during the last years. The focus of such applications is readability and fulfilling written language conventions, and therefore spoken language features such as false starts, mispronunciations and disfluencies are automatically corrected or left out. When studying spoken language and interpreting in particular, these spoken features are especially interesting and should be included in the transcripts. Therefore, the transcription process for such corpus material is still a process that requires substantial manual effort. A transcription of a spoken text can never fully mirror the acoustic signal (Bernardini et al., 2018). Depending on the research focus, linguistic, paralinguistic and extra-linguistic data will be included in greater or lesser detail. As our main research focus is on linguistic phenomena while ensuring comparability to other corpora of the EPIC family of corpora, we follow the transcription guidelines used for the compilation of EPICG, described in (Bernardini et al., 2018): As for the linguistic level, data were transcribed orthographically following the EU Interinstitutional Style Guide[1]. In order to align source and target speeches on sentence level, sentence boundaries have to be included. EPICG uses EXMARaLDA (Schmidt and Wörner, 2014) to time-align source and target. Sentence annotation is therefore not described in the guidelines. As our aim is also to analyse target language features that might be triggered by a source language expression, we need sentence-equivalent segments in order to perform such an alignment. Here we opt for a similar approach as in EPIC: Sentence boundaries are detected using syntactic information as well as speakers intonation (annotated with ↲). In spoken language, many main clauses are connected with coordinating conjunctions, leading to long parataxis. In order to get better alignment results during the next corpus compilation steps, we opt for the smallest possible sentence segment, splitting several coordinated clauses into individual segments such as in the example below, taken from the SI DE EN subcorpus:

- it's obvious for a country ruled by the state of law ↲
- and I think it's a great shame that we have to talk about this / as if it weren't obvious / ↲
- and therefore / the fact that we're saying / there should be a ceasefire / stop this weaponry / ↲
- and we should be able to do everything possible / in a humanitarian way to cope with this conflict / ↲

As for the paralinguistic level, filled and silent pauses were transcribed, truncated words and mispronunciations as well as possible ambiguities, following the EPICG guidelines in (Bernardini et al., 2018). Table 1 gives an overview of the transcription conventions used for the paralinguistic level. All transcriptions, revisions and segmentation of new and existing transcripts were carried out by one transcriber and validated by a second linguist.

Feature	Transcription
Silent pause	/
Filled pause	euh, hm, hum
Mid-word pause	spea/ euh ker [speaker]
Rising intonation	[?]
Non-verbalized noise	[noise], [breath]
Non-standard pronunciation	report [repo:rt]
Inaudible segment	[inaudible]
Mispronunciation	plemary [plenary]
Truncated word	propo/
Ambiguity	they [?there]
Sentence-equivalent unit	↲

Table 1: EPIC UdS transcription for interactional and non-verbal acoustic features based on EPICG (with minor modifications)

For more fine grained analysis purposes, metadata have to be stored as structural attributes. In accordance with EPIC and EPICG (Bernardini et al., 2018), Table 2 gives an overview of information on extra-linguistic level collected for EPIC UdS [2].

As in the existing EPIC corpora, source text delivery type was assigned depending on whether the source speaker could be seen reading a script (read) or not (impromptu), or switching between the two modes (mixed). Such extra-linguistic information is especially relevant for comparable corpus studies: Przybyl et al. (2022) show that interpreted speech compared to original spoken production in the EP is characterised by more spoken language features. In order to see whether this is a true interpreting effect or rather the result of some originals being written to be read out, studies need to take into account source speech delivery type.

[1] http://publications.europa.eu/code/en/en-300000.htm

[2] Speech delivery rate in accordance with classification in EPIC

Speaker related	Name Gender Nationality (operationalised for determining native language)
Interpreter related	Gender Native language
Speech event related	General topic (as indicated by EP) Title of debate (as indicated by EP) Date Speech length in words Speech length in seconds Speech delivery rate: in words per minute (w/m) Speech delivery rate: *slow* ≤130w/m; medium = 131-160w/m; *high* ≥161w/m Source text delivery type (read, impromptu, mixed)

Table 2: EPIC UdS metadata

3.3. Annotation

Depending on research focus, it is sensible to ignore some of the features annotated and transcribed in the first place: if we are looking at syntactic complexity via automatic dependency parsed output, it makes sense to remove filled and silent pauses as well as truncated words in order to get better quality parsing results. For studies of disfluencies and possible triggers, fine grained annotation thereof are required. These, however, are a challenge for many tagging tools. Often used taggers such as Treetagger produce erroneous output for spoken language features. Some of them can fairly easy be corrected (e.g. for filled pauses), other such as truncated words belonging to different word classes are more difficult to handle.

4. Corpus Variants

We provide several EPIC UdS corpus variants that can be used for different research purposes. The variants differ concerning the annotations used as well as spoken language features included. All corpus variants are accessible via https://corpora.clarin-d.uni-saarland.de/cqpweb/.

4.1. Comparable Subcorpora

EPIC UdS V2 is the comparable variant of the corpus. It can be used, for instance, for monolingual studies comparing the production of original speakers with that of interpreters in the same language. Table 3 gives an overview of EPIC UdS V2's subcorpora and their respective size.

Besides lemmatisation and POS-information as annotation layers (UPOS as well as fine grained language specific POS: e.g. STTS for German, UPENN for English), universal dependency relations are also encoded including dependency relation to the head (deprel) and head of the current word (head).

For this, spaCy NLP tools (version 2.3.4) (Honnibal and Montani, 2017) and the corresponding language models (de_core_news_lg-2.3.0, en_core_web_lg-2.3.1, es_core_news_lg-2.3.1) were used to process the sentence annotated transcripts. A randomly selected subset of 50 sentences for English and German was manually evaluated by two independent evaluators in order to assess the percentage of words with correct head (UAS: Unlabeled attachment score) and of words with correct head and label (LAS: Labeled attachment score). Table 4 gives an overview of the evaluated parsing accuracy for EPIC UdS V2 compared to stated parsing accuracy for the spaCy language models used. Due to spoken language features such as filled pauses, false starts and unfinished sentences present in the transcripts, parsing accuracy for both languages are well below the officially stated parsing accuracy. As simultaneous interpreting tends to use more spoken language features than the original speeches in the corpus (Przybyl et al., 2022), we expect accuracy results to be lower for SI compared to ORG. In order to increase parsing accuracy, we removed some spoken language features and used Stanza language models (v1.2.3) (Qi et al., 2020) in a further corpus variant (EPIC-UdS V3). With these measures we were able to increase accuracy scores to 85.94 (LAS) and 89.08 (UAS) for English and 85.78 (LAS) and 91.03 (UAS) for German. [3]

	Tokens	Sentences
EPIC UdS EN		
ORG EN	67,526	3,622
SI DE EN	58,503	3,623
SI ES EN	54,630	3,076
EPIC UdS DE		
ORG DE	56,488	3,409
SI EN DE	57,532	4,076
EPIC UdS ES		
ORG ES	53,947	2,537

Table 3: Corpus overview EPIC UdS V2

[3] Accuracy scores have only been evaluated for the comparable datasets EN and DE.

	LAS	UAS
stated spaCy accuracy scores EN	90.28	92.09
EPIC UdS EN parsing accuracy	78.73	86.42
stated spaCy accuracy scores DE	91.15	92.99
EPIC UdS DE parsing accuracy	69.74	76.64

Table 4: Parsing accuracy for spaCy language models and EPIC UdS V2

This corpus variant is available for download at http://hdl.handle.net/21.11119/0000-0008-F519-8.

4.2. Parallel Corpora

For the parallel and aligned variants of EPIC UdS we used Intertext to automatically align and manually correct source and target sentences. Two parallel corpus variants are available – EPIC UdS aligned/parsed and EPIC UdS aligned/POS – with different focus: The first aligned corpus variant includes dependency parsing. In order to achieve better parsing quality, we remove truncated words, filled and silent pauses etc. in this corpus variant. For comparability of source and target parses, the parser used in the above-mentioned V2 variant of the corpus is not suitable as spaCy uses language specific dependency tags. In the aligned and parsed variants, we therefore opt for the Stanford NLP Python Library Stanza, so that dependency relations can be compared across languages. The second aligned variant is lemmatised and POS-tagged. Here we include some spoken features (e.g., filled pauses) and if necessary, manually correct their incorrect POS assignment, in order to pursue studies on disfluencies.

These aligned corpus variants are currently being finalised for publication.

5. Applications

5.1. Comparable Corpora

We use the comparable corpus variant EPIC UdS V2 and the comparable written dataset Europarl UdS (Karakanta et al., 2018) in order to detect translationese/interpretese features. By applying a divergence measure (here: Kullback-Leibler Divergence) on probability distributions over words obtained by n-gram models, we compare translation and interpreting including comparison to their written/spoken originals (Przybyl et al., 2022). The study confirms the overall trend of written vs. spoken mode being strongly reflected in translation and interpreting output. Specifically for interpreting, we observe a higher degree of orality (e.g. expressed by such distinctive items as filled pauses, discourse markers, deictics and intensifiers) compared to spoken originals (Figure 1), which confirms previous observations made by Shlesinger and Ordan (2012).

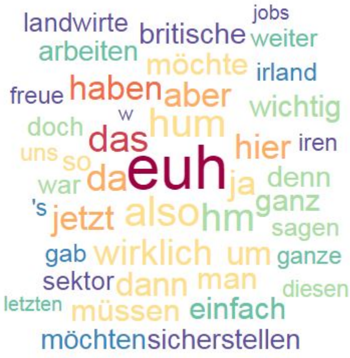

Figure 1: Distinctive features for interpreting compared to spoken originals: Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size (Przybyl et al., 2022).

The rich linguistic annotations of EPIC UdS V2, especially dependency parsing, are used to study syntactic complexity in interpreting, particularly dependency length minimisation (Przybyl and Teich, 2021). Using the annotated dependency relations, dependency distance of head and dependent were calculated for each word in the corpus. Syntactic complexity indicators such as average dependency length, maximum dependency length and number of adjacent dependencies for the English and German comparable corpora show a reduction of syntactic complexity in interpreted speech compared to originals in the same language. Especially long dependency relations occur less frequently in interpreting (cf. Figure 2). However, the trend of dependency length minimisation in interpreting can mainly be related to shorter sentences in interpreting than originals overall and can not be observed when comparing sentence of the same length in English.

5.2. Parallel Corpora

Recently, multilingual embeddings have proven to be very useful for multilingual computational modelling. Although, interpreting corpora are small in size, they can also be used for this approach, as it was shown for the parallel variants of our corpora. Both Bizzoni and Teich (2019) and Lapshinova-Koltunski et al. (2021b), compared bilingual word embeddings in translation and interpreting. In these studies, the aligned sentences were used to train a standard skip n-gram Word2Vec model and to create bilingual neural semantic spaces for translation and interpreting. Such spaces contain words with similar contexts, with context of each word being words in both languages of the aligned sentences in which a particular word occurs. The spaces of the same words in translation and interpreting were then analysed and compared. The main idea of this method is that words with a consistent translation share similar contexts with their equivalents and fall in close proximity. The imbalance in size between translation and interpreting corpora does not pose a problem for the comparison task, as it was shown in (Bizzoni and Teich, 2019, p. 5).

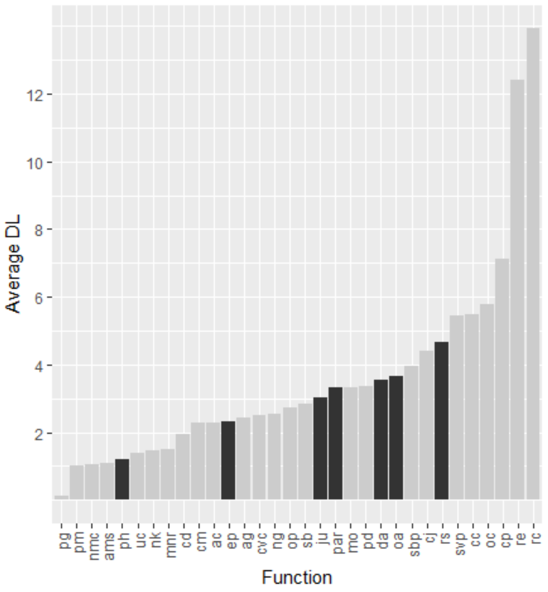

Figure 2: Dependency relations and their average dependency length for ORG DE vs SI DE EN). Black denotes more frequent use in original, grey marks more frequent use in simultaneous interpreting.

Bilingual embeddings trained on our corpora can be used for the analysis of various linguistic aspects. For instance, Bizzoni and Teich (2019) analyse general lexical differences between interpreting and translation. They found that formulaic or highly predictable words are always translated with the same equivalent in interpreting. Besides that, semantic spaces of interpreting seem to contain more domain-specific and technical terminology instead of more stylistic features, e.g. interpersonal and textual expressions or adverbs. At the same time, (Lapshinova-Koltunski et al., 2021b) focus on the level of discourse, comparing neural semantic spaces of discourse connectives in translation and interpreting. Apart from general differences between semantic spaces of interpreting and translation that were also reported by Bizzoni and Teich (2019), this study observed the following differences: Interpreting use more general connectives to mark logical relations, which is in line with the existing observations about differences between speech and writing (Crible and Cuenca, 2017). Besides that, interpreting exhibits reduced variability in connectives (fewer types) which leads to stronger clusters (i.e. cosine similarity is higher) if compared to translated texts (cf. Table 5). The fact that semantic spaces of many connectives in interpreted texts do not contain any equivalents point to implicitation effects also discussed in the literature about interpreting. In this way, interpreting shows more implicitation than translation. Apart from these general tendencies, we also discovered that cognitive complexity of relations has impact on the resulting semantic spaces in translation with cognitively more complex relations like concession being reflected in the translation space whereas cognitively simpler relations such as expansion and contingency more often being left out (as also reported by Hoek et al. (2017). However, this effect can not be seen in the interpreting spaces.

	TR space	**SI space**
if	when .73; unless .66; though .51	wenn .87; dann .72;
as	(angesehen .53)	wie 0.57
secondly	zweitens .76	zweitens .82

Table 5: Translation and interpreting spaces for *if, as* and *secondly* and the nearest neighbours with cosine similarity; semantically unrelated items in brackets (Lapshinova-Koltunski et al., 2021b).

6. Discussion and Future Work

We provide several variants of a European Parliament interpreting corpus for the language combinations English, German and Spanish that can be used for a vast range of applications and study purposes. This includes applications presented in this paper, but also specifically extends to cognitive aspects of the interpreting process traceable in the interpreting product as well as applications for interpreter training, applications in translation and interpreting research other than the ones mentioned in this paper or for scholars of conversation and discourse analysis.

Corpus-based approaches to interpreting process research to reflect cognitive load in interpreting have recently received more attention. Approaches include studying filled pauses triggered by numbers, lexical density, formulaicity and syntactic complexity (Plevoets and Defrancq, 2016; Defrancq and Plevoets, 2018; Plevoets and Defrancq, 2020) or cognates and low frequency words (Chmiel et al., 2021). Furthermore, phonetic features such as pitch are promising in mirroring cognitive load in simultaneous interpreting (Defrancq, 2021).

In order to allow for more fine grained analysis of phonetic features linked to cognitive load we also plan to align a subset of the data in Praat (Boersma and Weenink, 2001). This will enable us to study the interplay of interpreters lag, disfluencies and phonetic features. Furthermore, we are also working on improving the existing material by adding metadata that can not be easily obtained: We are currently including manual interpreter identification in order to exclude interpreter idiosyncrasies from the results.

EPIC UdS is a rich resource to be used on its own, in combination with written European Parliament corpora as shown in the applications discussed above, or combined with other interpreting corpora. With a focus on efficient use of existing resources we based our corpus on guidelines available for other corpora of the EPIC family, and therefore EPIC UdS can easily be used in combination with these resources in order to study more languages and language combinations than the limited amount of languages present in each of the individual EPIC family corpora.

7. Acknowledgements

This work is based on research funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1102 / Project-ID 232722074, project B7 - Modelling Human Translation with a Noisy Channel.

8. Bibliographical References

Bernardini, S., Ferraresi, A., and Miličević, M. (2016). From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective. *Target*, 28:61–86.

Bizzoni, Y. and Teich, E. (2019). Analyzing variation in translation through neural semantic spaces. In *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC) at RANLP-2019*, Varna, Bulgaria. ACL.

Chmiel, A., Janikowski, P., Danijel, K., Lijewska, A., Kajzer-Wietrzny, M., and Jakubowski, D. (2021). Lexical frequency modulates current cognitive load, but triggers no spillover effect in interpreting. In *Proceeding of the 3rd International Conference on Translation, Interpreting and Cognition*, Forli, November.

Crible, L. and Cuenca, M. J. (2017). Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8:149–166.

Defrancq, B. and Plevoets, K. (2018). Over-uh-load, filled pauses in compounds as a signal of cognitive load. In Mariachiara Russo, et al., editors, *Making Way in Corpus-based Interpreting Studies*, pages 43–64. Springer Singapore, Singapore.

Defrancq, B. (2018). The European Parliament as a discourse community: its role in comparable analyses of data drawn from parallel interpreting corpora. *The Interpreters' Newsletter*, 23:115–132.

Defrancq, B. (2021). The dark energy of simultaneous interpretation, November. Keynote speech at the 3rd International Conference on Translation, Interpreting and Cognition.

Marcello Federico, et al., editors. (2020). *Proceedings of the 17th International Conference on Spoken Language Translation*, Online, July. Association for Computational Linguistics.

Marcello Federico, et al., editors. (2021). *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, Bangkok, Thailand (online), August. Association for Computational Linguistics.

Ferraresi, A. and Miličević, M. (2017). Phraseological patterns in interpreting and translation. Similar or different? In G. De Sutter, et al., editors, *Empirical Translation Studies. New Methodological and Theoretical Traditions*, volume 300 of *Trends in Linguistics. Studies and Monographs [TiLSM]*, pages 157–182. Mouton de Gruyter.

Graves, A., Olaguíbel, M. P., and Pearson, C. (2021). Conference interpreting in the European Union institutions. In Michaela Albl-Mikasa et al., editors, *The Routledge Handbook of Conference Interpreting*, London. Routledge.

He, H., Boyd-Graber, J., and Daumé III, H. (2016). Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976. Association for Computational Linguistics.

Hoek, J., Zufferey, S., Evers-Vermeul, J., and Sanders, T. J. (2017). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131.

Kajzer-Wietrzny, M. (2015). Simplification in interpreting and translation. *Across Languages and Cultures*, 16(2):233–255.

Lapshinova-Koltunski, E., Bizzoni, Y., Przybyl, H., and Teich, E. (2021a). Found in translation/interpreting: combining data-driven and supervised methods to analyse cross-linguistically mediated communication. In *Proceedings of the Workshop on Modelling Translation: Translatology in the Digital Age (MoTra21)*, pages 82–90, online, May 31. Association for Computational Linguistics. co-located with NoDaLiDa-2021.

Lapshinova-Koltunski, E., Przybyl, H., and Bizzoni, Y. (2021b). Tracing variation in discourse connectives in translation and interpreting through neural semantic spaces. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse (CODI)*, pages 134–142, Punta Cana, Dominican Republic and Online, 10–11 November. co-located with EMNLP 2021.

Lapshinova-Koltunski, E. (2021). Analysing the dimension of mode in translation. In Mario Bisiada, editor, *Empirical Studies in Translation and Discourse*, Translation and Multilingual Natural Language Processing, pages 223–243. Language Science Press, Berlin.

Plevoets, K. and Defrancq, B. (2016). The effect of informational load on disfluencies in interpreting. a corpus-based regression analysis. *Translation and Interpreting Studies*, 11, 08.

Plevoets, K. and Defrancq, B. (2020). Imported load in simultaneous interpreting : an assessment. In Muñoz Martín, Ricardo and Halverson, Sandra L., editor, *Multilingual mediated communication and cognition*, The IATIS Yearbook, pages 18–43. Routledge.

Przybyl, H. and Teich, E. (2021). Dependency length minimization in simultaneous interpreting. In *Proceeding of the 3rd International Conference on Translation, Interpreting and Cognition*, Forli, November.

Przybyl, H., Karakanta, A., Menzel, K., and Teich, E. (2022). Exploring linguistic variation in mediated discourse: translation vs. interpreting. In Marta Kajzer-Wietrzny, et al., editors, *Empirical investigations into the forms of mediated discourse at the European Parliament*, Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Sandrelli, A. and Bendazzoli, C. (2005). Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). In *Proceedings from the Corpus Linguistics Conference Series*.

Shlesinger, M. and Ordan, N. (2012). More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target*, 24:43–60.

9. Language Resource References

Claudio Bendazzoli. (2010). *Il corpus DIRSI: creazione e sviluppo di un corpus elettronico per lo studio della direzionalità in interpretazione simultanea*.

Bernardini, Silvia and Ferraresi, Adriano and Russo, Mariachiara and Collard, Camille and Defrancq, Bart. (2018). *Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task*. Springer Singapore.

Boersma, Paul and Weenink, David. (2001). *PRAAT, a system for doing phonetics by computer*.

Chmiel, Agnieszka and Kajzer-Wietrzny, Marta and Koržinek, Danijel and Janikowski, Przemysław. (2021). *Cross-linguistic similarities in lexis: examining cognate activation through temporal and accuracy data from the Polish Interpreting Corpus (PINC)*.

Dayter, Daria. (2018). *Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN)*.

Defrancq, Bart and Plevoets, Koen and Magnifico, Cédric. (2015). *Connective Items in Interpreting and Translation: Where Do They Come From?* Springer International Publishing.

Doi, Kosuke and Sudoh, Katsuhito and Nakamura, Satoshi. (2021). *Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data*. Association for Computational Linguistics.

Honnibal, Matthew and Montani, Ines. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Kajzer-Wietrzny, Marta. (2012). *Interpreting universals and interpreting style*.

Karakanta, Alina and Vela, Mihaela and Teich, Elke. (2018). *Europarl-UdS: Preserving Metadata from Parliamentary Debates*. European Language Resources Association (ELRA).

Kunz, Kerstin and Stoll, Christoph and Klüber, Eva. (2021). *HeiCiC: A simultaneous interpreting corpus combining product and pre-process data*. Association for Computational Linguistics.

Macháček, Dominik and Žilinec, Matúš and Bojar, Ondřej. (2021). *ESIC – Europarl Simultaneous Interpreting Corpus*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3719, 1.0.

Pan, Jun. (2019). *The Chinese/English Political Interpreting Corpus (CEPIC): A New Electronic Resource for Translators and Interpreters*. Incoma Ltd., Shoumen, Bulgaria.

Qi, Peng and Zhang, Yuhao and Zhang, Yuhui and Bolton, Jason and Manning, Christopher D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*.

Russo, Mariachiara and Bendazzoli, Claudio and Monti, Cristina and Sandrelli, Annalisa and Baroni, Marco and Bernardini, Silvia and Mack,Gabriele and Piccioni, Lorenzo and Zanchetta, Eros and Ballardini, Elio and Mead, Peter. (2005). *European Parliament Interpreting Corpus*. ISLRN 716-168-855-843-2.

Thomas Schmidt and Kai Wörner. (2014). *EXMARaLDA*. Oxford University Press.

Temnikova, I. and Abdelali, A. and Hedaya, S. and Vogel, S. and Daher, A. Al. (2017). *Interpreting strategies annotation in the WAW corpus*.

Zhang, Ruiqing and Wang, Xiyang and Zhang, Chuanqiang and He, Zhongjun and Wu, Hua and Li, Zhi and Wang, Haifeng and Chen, Ying and Li, Qinfei. (2021). *BSTC: A Large-Scale Chinese-English Speech Translation Dataset*. Association for Computational Linguistics.