Tracing variation in discourse connectives in translation and interpreting through neural semantic spaces

Ekaterina Lapshinova-Koltunski **Heike Przybyl** **Yuri Bizzoni**
Saarland University, University Campus A.2.2, DE-66123 Saarbrücken
e.lapshinova@mx.uni-saarland.de
heike.przybyl@uni-saarland.de
yuri.bizzoni@uni-saarland.de

Abstract

In the present paper, we explore lexical contexts of discourse markers in translation and interpreting on the basis of word embeddings. Our special interest is on contextual variation of the same discourse markers in (written) translation vs. (simultaneous) interpreting. To explore this variation at the lexical level, we use a data-driven approach: we compare bilingual neural word embeddings trained on source-to-translation and source-to-interpreting aligned corpora. Our results show more variation of semantically related items in translation spaces vs. interpreting ones and a more consistent use of fewer connectives in interpreting. We also observe different trends with regard to the discourse relation types.

1 Introduction

This paper presents an explorative study of discourse connectives in cross-linguistically mediated communication. We compare written translation with spoken simultaneous interpreting in the domain of European Parliament discourse. We start from the perspective of *translationese*, an observation that translations share specific linguistic features distinguishing them from non-translated language. Although interpreting has recently received increased attention in computational approaches, there are not so many studies of *interpretese*.

In this work, we are aiming to understand the differences between translated and interpreted texts in terms of discourse connectives. Przybyl et al. (forthcoming) show that translations and interpreting transcripts differ significantly in the use of these linguistic units. For instance, there is a difference in the preference for the choice of connectives triggering the same relation. Specifically, the relation of contrast/concession is preferably expressed with the connective *however* in translation, whereas the use of *but* is characteristic of interpreting. This indicates that in marking logical relations, interpreters tend to prefer more general items over more specific ones. This is in line with the existing observations about differences between speech and writing (Crible and Cuenca, 2017): there are fewer but more polyfunctional discourse markers in speech than in writing.

To explore the differences between translated and interpreted texts, we follow the data-driven approach as in (Bizzoni and Teich, 2019) and use neural word embeddings (Word2Vec) to compare the bilingual semantic spaces from bilingual word embeddings built on aligned corpora. The resulting semantic spaces model the lexical choices of a specific translation. We train two bilingual distributional models on two comparable, aligned corpora (translation and interpreting) and compare the resulting semantic spaces to detect differences in the lexical patterns of discourse connectives impacted by translation mode (written vs. spoken). As stated by Bizzoni and Teich (2019) the existing constraints of interpreting (high cognitive load, time pressure) have impact on lexical choices, which is reflected in interpreting if compared to translation.

2 Related Work

Translated texts share linguistic characteristics which distinguish them from non-translated texts – the phenomenon of translationese (see Gellerstam, 1986; Baker, 1993; Toury, 1995). These differences can be traced in the distribution of various language patterns, i.e. linguistic features mostly organised in terms of more abstract categories (sometimes called translation universals or translation features) such as *normalisation* and *shining-through* (Teich, 2003), *simplification* (Toury, 1995), *convergence* (Laviosa, 2002) and *explicitation* (Olohan and Baker, 2000). The latter is often related to discourse connectives. For instance, Gumul (2006, 184) stated that explicitation in interpreting is related to adding discourse markers among other means of cohesive explicitness. At the same time, Shlesinger (1995) observed a reduction of cohe-

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 134–142
November 10–11, 2021. ©2021 Association for Computational Linguistics

sive ties in interpreting if compared to the source language input (implicitation). Kajzer-Wietrzny (2012) showed that there are differences between translation and interpreting in the usage of linking adverbials (with translation being more explicit).

The phenomena of explicitation and implicitation may also depend on the type of relations discourse connectives trigger: cognitively simple relations are more often left implicit than relations that are cognitively more complex (see Hoek et al., 2017). This is also confirmed in a recent study by Blumenthal-Dramé (2021) who showed that causal links are more expected than concessive ones and the processing of concessive sentences benefits more from the explicit marking than the processing of causal sentences (also pointing to cross-lingual differences between English and German).

Although we are not pursuing creation of a multilingual lexicon, our work is related to those dealing with mapping of discourse connectives (Stede et al., 2019; Bourgonje et al., 2018, 2017; Laali and Kosseim, 2017, 2014). Numerous studies analysed discourse connectives from a cross-lingual point of view using aligned texts (see Hoek and Zufferey, 2015; Zufferey and Cartoni, 2014; Meyer and Webber, 2013; Cartoni et al., 2011; Meyer et al., 2011). In our exploratory study of the differences between translation and interpreting, we do not focus on the direct transfer of specific discourse connectives, but look into their bilingual lexical context. In this sense, our work relates to the study by Roth and Upadhyay (2019), who used cross-lingual embeddings and discourse connectives to analyse semantically related words in several languages.

We rely on the methodology proposed by Bizzoni and Teich (2019), which is related to other studies that attempt to use word embeddings for linguistic analysis, such as Dubossarsky et al. (2017); Fankhauser and Kupietz (2017); Bizzoni et al. (2019). Whereas Bizzoni and Teich (2019) analyse general variation in lexical choices in interpreting and translation, our focus is on discourse-related phenomena. Besides that, our work is related to those with a focus on domain-specific word embeddings (Zhang et al., 2019; Wang et al., 2018), those dealing with multilingual word embeddings for lexicon induction and mapping (Shi et al., 2019; Zhang et al., 2019; Artetxe et al., 2017, 2018) and those focusing on creating consistent spaces (Huang et al., 2018) in cross-lingual word analogy tasks (Ulčar et al., 2020; Brychcín et al., 2019).

3 Data and Methods

Our data includes officially published original speeches, as well as transcripts of the speeches delivered at the European Parliament (EP) aligned with their translations or interpretations, correspondingly. Both parallel subsets are parts of the Europarl-UdS (Karakanta et al., 2018) and EPIC-UdS (Przybyl et al., forthcoming) corpora, with English as source and German as target. Both corpora are strictly comparable in terms of register, as they contain European Parliament speeches. The spoken part (EPIC-UdS) are true transcripts of the spoken utterances by members of the EP and interpeters, including spoken language features such as false starts, filled pauses and unfinished sentences. The basis of written dataset (Europarl-UdS) is also a spoken event in the European Parliament, however modified to respect written language characteristics. Overall, we have 130,000 sentence pairs in the translated data and 3,397 sentence pairs for the interpreted data. Refer to Table 1 for an overview of the size in tokens.

written	**token**	**spoken**	**token**
WR	9,654,581	SP	66,226
TR	8,954,825	SI	57,622

Table 1: Dataset used for written (Europarl-UdS) and spoken (EPIC-UdS). WR=English written original, TR=translation into German, SP=English spoken original, SI=simultaneous interpreting into German.

This paper uses the approach described by Bizzoni and Teich (2019): neural word embeddings (Word2Vec) are used to compare source-to-translation and source-to-interpreting lexical choices. The sentence aligned corpus data is used to create reshuffled bilingual pseudo-sentences. Subsequently, a standard skipgram Word2Vec model is trained on these sentences to create translation and interpreting spaces. The main idea of this method is that words with a consistent translation in an aligned corpus will share similar contexts with their translation, and will result in close proximity in distributional spaces. This is a mechanism similar to that used in standard distributional semantics, where words having similar contexts are closer in space, but it is applied to bilingual aligned contexts instead of standard monolingual texts.

For example, if a source word A is always translated with a target word B, and B is only used to translate A, such words will appear only together and their bilingual context will be identical. As an effect, their distributional vectors will be extremely similar and the two words will be very close neighbours in a semantic space. The more a translation deviates from this type of systematicity, the less two words will result close in the semantic space.

For the words without consistent translation, different configurations in translation space exist. If a word is ambiguous, it can be close to the variants of its translations in the space, but the similarity would be lower (reflected in the cosine similarity score). If a word is very hard or even impossible to translate with one term, there will be no translations close to the word, which does not show a high similarity to its neighbours. While this method, being based on Word2Vec, is sensitive to frequency effects, Bizzoni and Teich (2019) show that it is robust enough to obtain meaningful results from corpora few thousands sentences long (see Bizzoni and Teich, 2019, for more details and examples).

Our list of connectives is restricted to 14 items that appear to be most frequent in both written and spoken data at hand.[1] The connectives can be grouped according to their senses defined in PDTB-3 (Webber et al., 2019, p. 17) as temporal (*finally, first, firstly, secondly*), contingency (*as, because, if, so, why*), comparison (*but, however, yet*) and expansion (*also, that*). Some of them are ambiguous and may trigger different relations, e.g. the connective *as* is ambiguous and may express all the four meaning relations, although preference is given for temporal and contingency (Webber et al., 2019, p. 56). The connective *finally* is ambiguous between temporal meaning and the meaning of expansion, whereas the connective *if* is conditional, but may also express comparison.

The resulting semantic spaces of connectives are analyse in the following way: 10 nearest neighbours of the 14 discourse connectives are analysed for semantic relation to the connective – if an item expresses the same discourse function, i.e. source or target language synonym (e.g. *however – but – zwar*), can be used for paraphrasing the function (e.g. *why – for this reason, finally – to sum up*) or items that express other logical relations via discourse connective (e.g. *that – ob*), we consider them as semantically related.

We have a number of assumptions about the resulting semantic spaces. First of all, following the findings by Bizzoni and Teich (2019), we expect that the connectives translated consistently with the same target language equivalent will be very close to each other in the resulting semantic space. Besides that, we know that interpreters tend to use fewer connectives than translators. So we expect less variation in interpreting than translation, and consequently, fewer but closer semantic neighbours in interpreting than in translation. At the same time, as there should be more implicitation in interpreting than in translation, we expect more target language equivalents in the semantic spaces of translation than interpreting. Furthermore, we expect variation in the resulting semantic spaces with regard to the discourse relation a connective expresses.

4 Analysis and Observations

We qualitatively analyse the differences between resulting translation and interpreting spaces (TR and SI spaces in the following). For the sake of space, we will report on the most remarkable cases only. The whole list of spaces for both translation and interpreting is given in the Appendix.

Overall observations The TR spaces display far greater semantic proximity with synonyms or other semantically related items than the SI spaces. The interpreting space for 5 out of 14 discourse connectives under analysis (*also, but, however, so, yet*) shows no related item within the 10 nearest neighbours. The maximum amount of semantically related items in the list of 10 nearest neighbours in the interpreting space is 3, whereas up to 10 out of 10 nearest neighbours in the translation space are semantically linked to the connective studied.

Content of TR and SI spaces Translation spaces of some connectives (*but, however, yet, finally*) display many near synonyms in English or equivalents in German, however linked with lower cosine similarity score. By contrast, their interpreting spaces contain either few or no semantically related items. One of such examples is the discourse connective *but*, that we illustrate in Table 2.

Notably, this discourse connective is frequent in both translated and interpreted texts of our data, being even more frequent in interpreting than in

[1] We excluded some of the frequent connectives due to their high ambiguity (*and, or*). We also considered two discourse particles *well* and *now*, but excluded them from the current analysis, as they are more typical of spoken language, i.e. interpreting.

TR space	SI space
however .86; whilst .7; while .7; yet .66; nevertheless .62; though .57; although .53; nonetheless .48; zwar .48; (bilden .41)	(ausschluss .57); (vorrecht .56); (kein .54); (forum .5); (exercising .5); (aktualisiert .5); (abtreibungsrecht .49); (wahrheitsfindung .49); (scheint .48); (institutionelle .47)

Table 2: Translation and interpreting spaces for *but* and 10 nearest neighbours with cosine similarity; semantically unrelated items in brackets.

translation if compared to the other two discourse connectives of comparison under analysis (*however* and *yet*). However, the SI space does not contain any semantically related English words or German equivalents, while the TR space does so (e.g. the synonym *however* or the translation equivalent *zwar*). We assume that no matching equivalents within the nearest neighbours in interpreting confirms the general implicitation trend as discussed in the literature about interpreting: due to the time pressure and high cognitive load, interpreters tend to omit discourse connectives used in the source.

Scores in TR and SI spaces In general, very few semantically related items occur in the interpreting space. However, if there is a synonym or translation equivalent in the SI cluster, cosine similarity is generally higher in the interpreting space than for equivalent items in the translation space, as seen in Table 3 for the connectives *if, as* and *secondly*.

	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 3: Translation and interpreting spaces for *if, as* and *secondly* and the nearest neighbours with cosine similarity; semantically unrelated items in brackets.

This can be explained by the general tendency of interpreting to frequently use a smaller repository of discourse connectives. This reduced variation in interpreting leads to stronger clusters within the semantic space. For instance, *if* is consistently interpreted with the target language equivalent *wenn* and is frequently used in proximity to *dann*. Cosine similarity scores for *wenn* in the SI space with .87 is higher than the highest nearest neighbour cosine similarity score in the TR space (see Table 3). If interpreters add connectives in the German target and the English sources do not contain their triggers (see Defrancq, Bart, 2016), the systematicity of the translation is reduced and the connectives' similarity to their triggers is lower; thus, they might not appear at all in the semantic clusters we analyse.

Connective ambiguity Connectives *so* and *that* seem not to cluster well with semantically related words in either translation or interpreting space. This can be explained by their multiple functions: they not only serve as discourse connective, but also as other discourse elements, e.g. *that* can be used as demonstrative reference, whereas *so* can express clausal substitution or also be used as a modifying adverb or an intensifier. With the word embedding approach applied on the raw data (not annotated for true discourse connectives), we cannot distinguish between these functions.

Observations on discourse senses We also observe variation in the patterns for various types of discourse relations. For instance, connectives triggering the relation of comparison (*but, however* and *yet*), temporal connectives (*finally, first, firstly* and *secondly*) and the conditional *if* have almost always equivalents in the translation spaces. This confirms the dependency of the implicitation/explicitation process on the type of relation, as shown by Hoek et al. (2017) and Blumenthal-Dramé (2021), see Section 2 above. Cognitively more complex relations (concession, which is grouped within comparison and condition) cannot be easily left out, and have therefore almost always equivalents in translation spaces. At the same time, cognitively simple relations (expansion and contingency) do not necessarily do so. However, this observation is true for translation only and does not apply for the interpreting data. An exception is the conditional *if.* It clusters with its equivalents in both translation and interpreting space, however showing semantically related connectives in English in the translation space only (*when, unless*), see Table 3 and the whole spaces in the Appendix.

5 Conclusion and Future Work

We used neural semantic spaces to observe differences between discourse connectives in translation and interpreting. Generally, we observe similar trends reported by Bizzoni and Teich (2019) for the lexical differences between interpreting and translation. Our results are in line with implication and explicitation trends in translation. They confirm the assumption that interpreting shows more implicitation than translation. It was also interesting to see that cognitive complexity of relations also has impact on the resulting semantic spaces in translation, but has a different effect on the interpreting spaces.

In future, we plan to further investigate these differences using a wider range of connectives expressing different discourse relations. We would also like to systematically compare our results with the original aligned corpus, to provide a convincing qualitative test of the trends we have observed through our semantic spaces.

Also, Word2Vec represents one of the most efficient methods to produce a word's compact distributional profile, but it is not the only one. It could be interesting to compare the results of Word2Vec with other state of the art, non-contextualized[2] word embeddings when applied to the same corpus.

Moreover, our results also confirm the tendency of interpreting to show less variation in terms of the range of discourse connectives. Here, we would like to extend our work and include more discourse connectives. Finally, we plan to experiment with disambiguated data (include the cases of discourse relations only).

Acknowledgement

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.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In G. Francis Baker M. and E. Tognini-Bonelli, editors, *Text and Technology: in Honour of John Sinclair*, pages 233–250. Benjamins, Amsterdam.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Katrin Menzel, Pauline Krielke, and Elke Teich. 2019. Grammar and meaning: Analysing the topology of diachronic word embeddings. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 175–185, Florence, Italy. Association for Computational Linguistics.

Yuri Bizzoni and Elke Teich. 2019. Analyzing variation in translation through neural semantic spaces. *Special topic: Neural Networks for Building and Using Comparable Corpora, Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria.*

Alice Blumenthal-Dramé. 2021. The online processing of causal and concessive relations: Comparing native speakers of english and german. *Discourse Processes*, 0(0):1–20.

Peter Bourgonje, Yulia Grishina, and Manfred Stede. 2017. Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics*, Rome, Italy.

Peter Bourgonje, Jet Hoek, Jacqueline Evers-Vermeul, Gisela Redeker, Ted Sanders, and Manfred Stede. 2018. Constructing a lexicon of dutch discourse connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175.

Tomáš Brychcín, Stephen Taylor, and Lukáš Svoboda. 2019. Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Systems with Applications*, 135:287–295.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How comparable are parallel corpora? measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 78–86, Portland, Oregon. Association for Computational Linguistics.

Ludivine Crible and Maria Josep Cuenca. 2017. Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8:149–166.

[2] Recent contextualized embeddings such as those produced by BERT and similar Transformers aim at capturing the a word's sense in context. Embeddings of this sort function like nuanced token vectors that change with every instance of a type. While this feature makes them excellent tools to build representations of larger semantic units, they could be a weakness for our experiment, since we aim at modelling the distributional profile of a word in a corpus.

Defrancq, Bart. 2016. Well, interpreters… a corpus-based study of a pragmatic particle used by simultaneous interpreters. In Corpas Pastor, Gloria and Seghiri Dominguez, Miriam, editor, *Corpus-based approaches to translation and interpreting : from theory to applications*, volume 106 of *Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation*, pages 105–128. Peter Lang.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Peter Fankhauser and Marc Kupietz. 2017. Visualizinglanguage change in a corpus of contemporary German. In *Proceedings of the 9th International Corpus Linguistics Conference*, University of Birmingham.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Ewa Gumul. 2006. Explicitation in simultaneous interpreting: A strategy or a by-product of language mediation? *Across Languages and Cultures. A Multidisciplinary Journal for Translation and Interpreting Studies*, 7:171–190.

Jet Hoek and Sandrine Zufferey. 2015. Factors influencing the implicitation of discourse relations across languages. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.

Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2017. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131.

Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Multi-lingual common semantic space construction via cluster-consistent word embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 250–260, Brussels, Belgium. Association for Computational Linguistics.

Marta Kajzer-Wietrzny. 2012. *Interpreting universals and interpreting style*. Ph.D. thesis, Uniwersytet im. Adama Mickiewicza, Poznan, Poland. Unpublished PhD thesis.

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. Europarl-UdS: Preserving metadata from parliamentary debates. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 610–619, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Majid Laali and Leila Kosseim. 2017. Automatic mapping of French discourse connectives to PDTB discourse relations. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–6, Saarbrücken, Germany. Association for Computational Linguistics.

Sara Laviosa. 2002. *Corpus-based Translation Studies, Theory, Findings, Application*. Rodopi, Amsterdam.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Proceedings of the SIGDIAL 2011 Conference*, pages 194–203, Portland, Oregon. Association for Computational Linguistics.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Maeve Olohan and Mona Baker. 2000. Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, 1:141–158.

Heike Przybyl, Alina Karakanta, Katrin Menzel, and Elke Teich. forthcoming. Exploring linguistic variation in mediated discourse: translation vs. interpreting. In Marta Kajzer-Wietrzny, Silvia Bernardini, Adriano Ferraresi, and Ilmari Ivaska, editors, *Empirical investigations into the forms of mediated discourse at the European Parliament*, Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Michael Roth and Shyam Upadhyay. 2019. Combining discourse markers and cross-lingual embeddings for synonym–antonym classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3899–3905, Minneapolis, Minnesota. Association for Computational Linguistics.

Weijia Shi, Muhao Chen, Yingtao Tian, and Kai-Wei Chang. 2019. Learning bilingual word embeddings using lexical definitions. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 142–147, Florence, Italy. Association for Computational Linguistics.

Miriam Shlesinger. 1995. Shifts in cohesion in simultaneous interpreting. *The Translator*, 1:193–214.

Manfred Stede, Amália Mendes, and Tatjana Scheffler. 2019. Connective-lex: A web-based multilingual lexical resource for connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*, 24.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Gideon Toury. 1995. *Descriptive Translation Studies – and Beyond*. John Benjamins, Amsterdam.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. *The Penn Discourse Treebank 3.0 Annotation Manual*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scintific Data*, 6.

Sandrine Zufferey and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target*, 26(3):361–384.

A Appendix

	TR space	SI space
temporal		
finally	abschließend .67; conclusion .59; lastly .55; (reiterate .53); schluss .5; (bemerken .49); abschluss .45; schließlich .42; conclude .42; (hervorheben .41)	sum .71; (unterbrechen .7); (lehrt .69); (gollnisch .69); (karte .68); (teach .68); (fish .66); (gollnish .66); (subject .65); (heben .65)
first	zweiten .61; second .58; firstly .57; ersten .51; zweite .49; mal .48; erstmals .46; zunächst .45; erste .44; (reading .43)	zunächst .76; (verbesserte .74); (französichen .74); (november .73); ersten .73; (mandats .72); erste .72; (wahl .71); (ehre .71); (vortragen .7)
firstly	erstens .75; secondly .72; zweitens .66; thirdly .6; first .57; drittens .56; (pfeiler .51); zunächst 0,50; (pillar .50); second .48	(achtzehn .87); endlich .83; (eighteen .83); (ausreichende .82); (gefallen .82); (umweltausschusses .81); (gordon .8); (mitkollaborateure .8); (erfasst .8); (mcavan 0,8)
secondly	zweitens .76; erstens .75; firstly .72; drittens .68; thirdly .67; fourthly .55; second .55; zweite .45; (pillar .44); viertens .44	zweitens .82; (procurement .77); (smes .74); (calculations .72); (roughly .72); (berechnungen .7); (fernen .7); (billionen .7); (fünfhundertachtzig .7); (edf .7)
contingency		
as	(angesehen .53); (erstes .52); (bezeichnen .52); (bezeichnet .51); (betrachtet .48); (than .48); (how .47); (erweisen .45); (gut .45); (insofern .45)	wie .57; (beginnt .57); (begins .56); (wirklichen .56); (arbeitsdokument .55); european .53; presently .53; (getreten .52); (far .52); (worse .52)
because	since .38; (geschweige .37); (owing .28); as .24; (moreira .24); (louth .24); (kleinlich .23); (timed .23); (improper .23); (vilify .23)	(dumping .62); (rückstände .6); (hoher .6); (güte .59); (ninety .57); denn .56; attributed .54; weil .53; (awful .53); (food .52)
if	when .73; unless .66; though .51; whenever .49, albeit .48; (ansieht .47); (durchkommt .47); (bedenkt .46); dann .44; whether .42	wenn .87; dann .72; (überhaupt .7); (nachzukommen .65); (cannot .64); (ohne .64); (prioritäten .64); (fährt .64); (without .62); (properly .62)
so	(genannten .4); (genannte .4); (möglich .39); (called .38); (schnell .37); therefore .36; why .35; (weitermachen .32); (getan .28); (quickly .28)	(oben .57); (schuhe .56); (represent .56); (unfortunate .56); (fight .55); (outer .55); (gut .55); (darzustellen .55); (vertrete .54); (sought .53)
why	reason .61; warum .59; weshalb .59; grund .54; explain .53; gründe .53; wieso .52; reasons .40; deshalb .39; therefore .38	(mainstream .56); (legislativpaket .56); (umfasst .56); (backed .55); (letztes .54); deshalb .53; (started .52); (weaken .51); (doubled .51); (gebeten .5)
comparison		
but	however .86; whilst .7; while .7; yet .66; nevertheless .62; though .57; although .53; nonetheless .48; zwar .48; (bilden .41)	(ausschluss .57); (vorrrecht .56); (kein .54); (forum .5); (exercising .5); (aktualisiert .5); (abtreibungsrecht .49); (wahrheitsfindung .49); (scheint .48); (institutionelle .47)

continued on next page

continued from previous page		
	TR space	**SI space**
however	but .86; nevertheless .73; yet .71; whilst .68; while .67; though .55; nonetheless .54; zwar .50; although .49; (coin .45)	(zweitausendzwölf .77); (informationssystem .77); (spiele .75); (achtunddreißig .74), (uk .73); (abtreibungsrecht .73); (paragraph .73); (characters .73); (zauberstab .72); (lewis .72)
yet	however .71; but .66; nevertheless .6; while .55; whilst .51; though .45; zwar .44; although .44, nonetheless .41; (stretches .39)	(destroy .64); (wussten .64); (produzierte .63); (pork .63); (kontrollierte .62); (zwang .61); (vernichten .61); (industrial .61); (throw .61); (wissenschaftler .60)
expansion		
also	furthermore .72; addition .66; similarly .64; equally .6; (indeed .57); moreover .57; sowohl .56; including .53; likewise .48; too .46	(sicherheit .54); (maßnahmen .51); (denk .51); (therapeuten .49); (gesundheitswesen .49); (repeal .49); (security .48); (fidschi .48); (interessen .48); (hum .48)
that	which .37; what .32; (tatsache .29); (assertion .29); (firmly .28); whatever .28; (richtig .27); (this .27); (klar .26); (sicherzustellen .26)	(verordnung .55); (daran .54); (fair .52); (veränderung .52); (study .52); (transparent .51); (incidentally .51); (offenheit .51); (spüren .51); (ob .5)

Table 4: Translation and interpreting spaces containing their 10 nearest neighbours with cosine similarity; semantically unrelated items in brackets.