Using Corpora in Contrastive and Translation Studies (7th edition)

Book of Abstracts

Adam Mickiewicz University, Poznań

10-12th July 2023

ADAM MICKIEWICZ
UNIVERSITY
POZNAŃ

Honorary Patronage of the Mayor of Poznań – Mr Jacek Jaśkowiak

Committees

Conference convenors

Marta Kajzer-Wietrzny

Agnieszka Chmiel

Organising committee

Marika Adamczyk-Żylińska

Iwona Mazur

Magdalena Perdek

Olga Witczak

Scientific committee

Silvia Bernardini

Łucja Biel

Mario Bisiada

Lynne Bowker

Sara Castagnoli

Daria Dayter

Gert DeSutter

Bart Defrancq

Adriano Ferraresi

Ana Frankenberg-Garcia

Łukasz Grabowski

Sylviane Granger

Ilmari Ivaska

Dorothy Kenny

Haidee Kotze

Karolina Krawczak

Kerstin Kunz

Natalie Kubler

Ekaterina Lapshinova-Koltunski

Marie-Aude Lefer

Rudy Loock

Sandra Halverson

Lorenzo Mastropierro

Tamara Mikolič Južnič

Maja Milicevic-Petrovic

Stella Neumann

Mariachiara Russo

Sonia Vandepitte

Bogusława Whyatt

Marion Winters

Acknowledgements

We would like to thank the partners of the UCCTS 2023 conference:

- The School of Languages and Literatures, Adam Mickiewicz University
- ID-UB, Exellence Initiative Research University Programme at Adam Mickiewicz University
- Lexical Computing
- Language Science Press
- Bloomsbury Publishing Plc

Advancing terminology through corpora: insights from Translation Studies

Łucja Biel

University of Warsaw, Poland

Terms are crucial components of special languages ISO (2019) — they are units of language, units of knowledge and units of communication (Cabré Castellví, 1999). They are fundamental building blocks of specialised discourses which activate rich knowledge structures. As a frequent source of difficulty in translation, they should naturally be of interest to translation scholars. Yet until recently corpus translation studies into specialised translation focused mainly on (stylistic) features of translated texts, the so called translation universals, largely ignoring terms. Corpora have been extensively and successfully used in terminography to study the behavior of individual terms for the purposes of preparing terminological resources. Yet researching terminological *patterns* systematically is admittedly technically and methodologically complex as they are more difficult to be queried.

Recent years have shown increased interest in terminology and attempts to overcome these obstacles. The interest has been fuelled by major theoretical and methodological developments in the field of Terminology itself. On the theoretical side, Terminology has diversified considerably, embracing cognitive, communicative and social aspects of the use of terms (Faber & Homme, 2022). On the methodological side, the advent of corpus linguistics and digital tools has facilitated term extraction and term analysis on an unprecedented scale, shifting attention to semasiological approaches and raising questions as to how we research terms.

The talk will explore how corpora enhance our understanding of terminological units and their behaviour in special languages and translation. I will illustrate the use of corpora for terminological research purposes with legal terms. One of the traditional areas of research includes translation techniques, with corpora allowing for scaling up the studies and their more nuanced contextualisation both as regards interlingual translation (Bajčić and Dobrić Basaneže (2021); Peruzzo (2019)) and intralingual translation (Biel & Doczekalska, 2020). Echoing the current interest in Terminology, translation scholars have started to intensively research variation triggered in translation (Vigier and Sánchez Ramos (2017); Prieto Ramos and Guzmán (2018); Guzmán and Prieto Ramos (2021)) or existing in source texts (Biel & Koźbiał, 2020). In particular, the use of parallel corpora makes it relatively easy to efficiently identify target-language equivalents of a source term, their variants and to measure the scale of intratextual and intertextual variation. Other important topics involve attempts to measure the degree of specialisation (Marín Pérez, 2016) and thematic hybridity of terms (Prieto Ramos & Cerutti, 2022). Methodologically, scholars have been experimenting with corpus designs, such as comparable and parallel corpora, genre-controlled and thematically-controlled corpora, and various techniques, such as n-grams (Biel & Doczekalska, 2020), binomials (Dobrić Basaneže (2018); Koźbiał (2020)); lexicometric analysis (Prieto Ramos & Morales Moreno, 2019), lexical profiles and statistical modelling (Roshdy, 2023).

The final part will present the project *Termness of terms: Anatomy of legal terms* which uses a mix of corpus linguistics, contrastive linguistics and language acquisition methods to analyse the complexity of legal terms in hybrid supranational (EU) and more traditional national (UK, IR) contexts via a prism of general language. Since popular automatic term extractors turned out to be very inaccurate in extracting terms from our legislative corpora due to their much less predictable and typical structure (term grammar), I extracted a special category of terms — defined terms — to study the complexity of and functions of their form in monolingual and multilingual settings. The fundamental question behind this study is what a (prototypical) term is and I believe that corpus research may validly contribute to answering this research question.

References

Bajčić, M., & Dobrić Basaneže, K. (2021, 2021/09/03). Considering foreignization and domestication in EU legal translation: a corpus-based study. *Perspectives, 29*(5), 706-721. https://doi.org/10.1080/0907676X.2020.1794016

Biel, Ł., & Doczekalska, A. (2020). How do supranational terms transfer into national legal systems?: A corpus-informed study of EU English terminology in consumer protection directives and UK, Irish and Maltese transposing acts. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 26*(2), 184-212. https://doi.org/https://doi.org/10.1075/term.00050.bie

Biel, Ł., & Koźbiał, D. (2020). How do translators handle (near-) synonymous legal terms? A mixed-genre parallel corpus study into the variation of EU English-Polish competition law terminology. *Estudios de Traducción 10*, 69-90. https://doi.org/https://dx.doi.org/10.5209/estr.68054

Cabré Castellví, M. T. (1999). *Terminology: Theory, methods and applications*. John Benjamins. https://www.jbe-platform.com/content/books/9789027298652

Dobrić Basaneže, K. (2018). Binomials in EU Competition Law. In S. Marino, Ł. Biel, M. Bajčić, & V. Sosoni (Eds.), *Language and Law: The Role of Language and Translation in EU Competition Law* (pp. 225-248). Springer International Publishing. https://doi.org/10.1007/978-3-319-90905-9_13

Faber, P., & Homme, M.-C. (Eds.). (2022). *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge*. John Benjamins. https://doi.org/https://doi.org/10.1075/tlrp.23.

Guzmán, D., & Prieto Ramos, F. (2021). Assessing legal terminological variation in institutional translation: The case of national court names in the human rights monitoring procedures of the United Nations. *Translation and Translanguaging in Multilingual Contexts, 7*(2), 224-247. https://doi.org/https://doi.org/10.1075/ttmc.00067.guz

ISO (2019). ISO 1087:2019 Terminology work and terminology science - Vocabulary. International Standardization Organization.

Koźbiał, D. (2020). Translation of Judgments: A Corpus Study of the Textual Fit of EU to Polish Judgments University of Warsaw]. Warsaw.

Marín Pérez, M. J. (2016). Measuring the degree of specialisation of sub-technical legal terms through corpus comparison: A domain-independent method. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 22*(1), 80-102. https://doi.org/https://doi.org/10.1075/term.22.1.04mar

Peruzzo, K. (2019). When international case-law meets national law: A corpus-based study on Italian system-bound loan words in ECtHR judgments. *Translation Spaces, 8*(1), 12-38. https://doi.org/https://doi.org/10.1075/ts.00011.per

Prieto Ramos, F., & Cerutti, G. (2022). Terminological hybridity in institutional legal translation: A corpus-driven analysis of key genres of EU and international law. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*. https://doi.org/https://doi.org/10.1075/term.21047.pri

Prieto Ramos, F., & Guzmán, D. (2018). Legal Terminology Consistency and Adequacy as Quality Indicators in Institutional Translation: A Mixed-Method Comparative Study. In F. Prieto Ramos (Ed.), *Institutional Translation for International Governance: Enhancing Quality in Multilingual Legal Communication* (pp. 81-101). Bloomsbury.

Prieto Ramos, F., & Morales Moreno, A. (2019). Terminological innovation and harmonization at international organizations: Can too many cooks spoil the broth? In I. Simonnæs & M. Kristiansen (Eds.), *Legal Translation. Current Issues and Challenges in Research, Methods and Applications* (pp. 87-110). Frank & Timme.

Roshdy, R. (2023). Translating Islamic Law: the Postcolonial Quest for Minority Representation Dublic City University].

Vigier, F. J., & Sánchez Ramos, M. d. M. (2017). Using parallel corpora to study the translation of legal system-bound terms: The case of names of English and Spanish courts. In R. Mitkov (Ed.), *Computational and corpus-based phraseology, Second International Conference, Europhras 2017, London, proceedings* (pp. 260-273). Springer.

Broadening the scope: the gravitational pull hypothesis in a usage-based theory of translation

Sandra L. Halverson

University of Adger, Norway

KEYWORDS: usage-based theory, gravitational pull, entrenchment – conventionalization model, linguistic knowledge

The gravitational pull hypothesis was introduced in Halverson (2003) as a potential explanation for differences in the distribution of some linguistic features in translated as opposed to non-translated language. More specifically, the hypothesis was put forward as a possible explanation for normalization/sanitization/conventionalization and generalization. The basic idea is that linguistic choice is impacted by the structure of linguistic knowledge (cp. Langacker, 1987) and that online structural characteristics impact translational choices in specific ways. The hypothesis was further refined in Halverson (2017).

In the period following the first articulation of the hypothesis a growing body of empirical research has provided the grounds for refining the underlying idea and situating it within a broader linguistic theory of translation, i.e., a usage-based theory (Halverson & Kotze, 2022). This talk starts by reviewing the current status of the hypothesis, later work of relevance to its main idea, and a subsequent body of work to test and refine it. This work includes studies done within the machine learning paradigm (Baroni & Bernardini, 2006; Volansky et al., 2015; Rabinovich et al., 2016) which suggest that the most successful discriminating features are frequency and 'interference' from the source text. In addition, a number of corpus-based studies, including multimethod studies, test the hypothesis on a number of language-pairs. This work also broadens the range of linguistic items used to test the hypothesis from the more predominantly lexical study in Halverson (2017) to the study of grammatical constructions in Hareide (2016), Marco (2021), and Lefer and de Sutter (2022) and to collocations in Ferraresi and Bernardini (2023). The most recent work also represents important methodological innovation.

Following this review, and based on the state-of knowledge emerging from it, the focus of the talk shifts to situating the most recent version of the hypothesis within a usage-based theory of translation. A usage-based theory of translation takes the usage-based paradigm in linguistics as its starting point and two current approaches are adopted for the purposes of this talk: Schmid's entrenchment-conventionalization model (2015, 2017, 2020) and Diessel's grammar network model (2019). In this part of the talk, the gravitational pull hypothesis is situated within this broader model, as part of a set of factors impacting linguistic choice.

The linguistic model is then coupled with a well-known socio-cognitive model of translation, Risku et al.'s (2013) sociocognitive model of translatorial cognition and action. This model provides more detail on the translational usage situation and the role of the various artifacts and actors involved in translation as we know it. The model allows us to situate translational decision-making in the technological and peopled domains in which it takes place and to study the direction of impact of human and technological factors. Finally, the relationships between this work and Kotze's constrained communication framework (2022) are also outlined. The synthesis of these approaches provides us with a theoretical framework that is sociocognitive in nature, broad in scope, and detailed enough at multiple levels to ground both empirical studies and continued theory development.

In the closing remarks, several critical questions are raised concerning the desired scope of a linguistic theory of translation, the utility of metaphorical constructs and a possible agenda for continued development in this area.

References

Baroni, M. & S. Bernardini. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing 21*(3), 259–274.

Diessel, H. (2019). *The grammar network*. Cambridge University Press.

Ferraresi, A. & S. Bernardini. (2023). Comparing collocations in translated and learner language. In search of a method. *International Journal of Learner Corpus Research 9*(1), 125–153.

Halverson, S. (2003). The cognitive basis of translation universals. *Target 15*(2), 197–241.

Halverson, S. (2017). Developing a cognitive semantic model: magnetism, gravitational pull, and questions of data and method. In G. de Sutter, M.-A. Lefer & I. Delaere (Eds.), *Empirical Translation Studies. New methods and theoretical traditions* (pp. 9-45). Mouton de Gruyter.

Halverson, S. & H. Kotze. (2022). Sociocognitive constructs in translation and interpreting studies (TIS): Do we really need concepts like norms and risk when we have a comprehensive usage-based theory of language? In S. L. Halverson & Á. Marín García (Eds.), *Contesting epistemologies in cognitive translation and interpreting* (pp. 51-79). Routledge. Hareide, L. (2016). The translation of formal source-language lacunas: An empirical study of the over-representation of target-language specific features and the unique items hypotheses. In M. Ji, L. Hareide, D. Li & M. P. Oakes (Eds.), *Corpus Methodologies Explained An empirical approach to translation studies* (pp. 137–187). Routledge.

Kotze, H. (2022). Translation as constrained communication. In S. Granger, & M.-A. Lefer (Eds.), *Extending the scope of corpus-based translation studies* (pp. 67–98). Bloomsbury.

Langaker, R. (1987). *Foundations of cognitive grammar. Volume 1.* Stanford University Press.

Lefer, M.-A. & G. de Sutter. (2022). Using the gravitational pull hypothesis to explain patterns in interpreting and translation: the case of concatenated nouns in mediated European Parliament discourse. In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska & S. Bernardini (Eds.), *Mediated discourse at the European Parliament: empirical investigations. Translation and Multilingual Natural Language Processing 19* (pp.133–159). Language Science Press.

Marco, J. (2021). Testing the Gravitational Pull Hypothesis on modal verbs expressing obligation and necessity in Catalan through the COVALT corpus. In M. Bisiada (Ed.), *Empirical studies in translation and discourse* (pp. 27-52). Language Science Press.

Rabinovich. E., S. Nisioi, N. Ordan & S. Wintner. (2016) On the similarities between native, non-native and translated texts. arXiv:1609.03204v1 [cs.CL] 11 Sep 2016.

Risku, H., F. Windhager & M. Apfelthaler. (2013). Extended translation. A sociocognitive research agenda. *Translation Spaces 2*(1), 151-182.

Schmid, H.-J. (2015). A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association 3*, 1–27.

Schmid, H.-J. (2017). A framework for understanding linguistic entrenchment and its psychological foundations. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning* (pp. 9–35). APA and Walter de Gruyter.

Schmid, H.-J. (2020). The dynamics of the linguistic system. Usage, conventionalization, and *entrenchment*. Oxford University Press.

Volansky, V., N. Ordan & S. Wintner. (2015). On the features of translationese. *Digital Scholarship in the Humanities* 30(1), 98–118.

Building Corpora for Contrastive, Translation and Interpreting Studies: A Journey from Unimodality to Multimodality

Sílvia Gabarró-López

Pompeu Fabra University, Spain & University of Namur, Belgium

KEYWORDS: spoken languages, visual signed languages, tactile signed languages, multimodality, corpora

For years, the study of spoken languages, on the basis of written and then also oral productions, was the only way to investigate the human language capacity. The numerous misconceptions that surrounded signed languages – i.e., the belief that they were pantomime, inferior communication systems which could convey limited meaning or a gestural transcoding of spoken languages – had excluded them from the research agenda. In 1960, the work of Stokoe showed that signed languages were not unstructured systems of gestures or pantomime used by deaf people, but fully-fledged languages with the same linguistic levels that can be found in any natural language.

This talk will be divided into four parts. In the first part, I will guide the audience through the history of signed language research, which is traditionally divided into three periods (Perniss, et al. 2007; Vermeerbergen & Nilsson, 2018). In the first period (1960 – 1980), scholars had to prove the linguistic status of signed languages by showing that they shared features with spoken languages. In the second period (1980 – 1990), the specificities of signed languages – i.e., iconicity, simultaneity and the use of space – and the differences with spoken languages were showcased. The third period started at the end of the 1990s. It was marked by an increasing interest in signed language typology and by the beginning of comparative research including gesture and spoken languages.

The second part of the talk will be devoted to the description of the reference signed language corpora. The advent of new technologies at the beginning of 2000s allowed the collection and storage of large datasets. The Auslan (Australian Sign Language) Corpus was the first of this kind (Johnston, 2010). This initiative was followed by other groups of scholars around the world. Existing signed language corpora are annotated to different extents, but most of them include ID-glosses – written labels in the ambient spoken language that are consistently used to identify signs throughout the corpus, regardless of their meaning in the context – and the translation into the written environmental spoken language (Johnston, 2010). Therefore, signed language corpora are not monolingual in the traditional sense, but they could be considered as some sort of unidirectional parallel corpora. Paradoxically though, they are not originally designed to investigate translations, but rather to document these endangered languages and to carry out corpus-based language descriptions (Hodge et al., 2019). Annotation protocols of signed language corpora vary from one corpus project to another, but they share some specificities in terms of content, making their data comparable.

In the third part of the talk, I will outline the changes that corpora have brought to signed language research and the different types of corpora that are being collected. The collection of signed language corpora has allowed the move from generative approaches to functional usage-based approaches, more representativity (more signed language users, more varied discourse genres), the growth of discourse studies, sociolinguistic studies, contrastive intra-modal and inter-modal studies, translation studies and the preservation of languages and cultures without a written tradition, among other things. At present, the collection of comparable signed and spoken language corpora is on the way (Hodge et al., 2019; Meurant et al., ongoing) as well as the collection of parallel corpora, containing signed and written/oral data (e.g., Nicodemus et al., 2017; Gabarró-López, 2018). Other types of corpora have also been collected such as tactile signed language corpora of deafblind signers (Mesch, 2016; Bono et al., 2018), and a tactile signed language corpus of deafblind signers from different signed languages and hearing interpreters (Raanes & Mesch, 2019). These corpora include first language users, but there are also corpora of learners (e.g.,

Schönstrom & Mesch, 2017; Leeson et al., 2020). The collection of all these datasets is allowing us to develop both monolingual and contrastive studies from different perspectives.

In the last part of the talk, I will present some challenges and future avenues related to the different types of signed language corpora. The first challenge faced by scholars is the difficulty of filming, recording and archiving data. Work needs to be done hand in hand with the community and ethical issues such as anonymization need to be addressed. Transcribing and segmenting data is not only a time-consuming task (which is mostly done manually), but it is also difficult to have it financed. Because of these challenges (and others), annotated data are still scarce (as compared to oral or written spoken corpora) and existing studies still rely on a limited number of signers/speakers. Yet, the development of these datasets and of contrastive studies between signed and spoken languages might have as significant an impact in the coming decades as the development of oral data has had on language knowledge since the 1980s.

References

Bono, M., Sakaida, R., Makino, R., Okada, T., Kikuchi, K., Cibulka, M., Willoughby, L., Iwasaki, S., & Fukishima, S. (2018). Tactile Japanese Sign Language and Finger Braille: An Example of Data Collection for Minority Languages in Japan. In *Proceedings of the 2018 LREC Conference*, N. Calzolari et al. (eds). http://lrec-conf.org/workshops/lrec2018/W1/pdf/18027_W1.pdf

Gabarró-López, S. (2018). *CorMILS: Pilot Multimodal Corpus of French – French Belgian Sign Language (LSFB) interpreters*. Stockholm: Department of Linguistics, Stockholm University, Namur: University of Namur.

Hodge, G., Sekine, K., Schembri, A. & Johnston, T. (2019). Comparing Signers and Speakers: Building a Directly Comparable Corpus of Auslan and Australian English. *Corpora* 14(1): 63 – 76.

Johnston, T. (2010). From Archive to Corpus: Transcription and Annotation in the Creation of Signed Language Corpora. *International Journal of Corpus Linguistics* 15(1): 106 – 131.

Leeson, L., Sheridan, S., Cannon, K., Murphy, T., Newman, H., & Veldheer, H. (2020). Hands in motion: Learning to fingerspell in Irish Sign Language. *Teanga* 11 (Special Issue 11): 120 – 141.

Meurant, L., Lepeut, A., Tavier, A., Vandenitte, S., Lombart, C., Gabarró-López, S. & Sinte, A. (Ongoing). *The Multimodal FRAPé Corpus: Towards Building a Comparable LSFB and Belgian French Corpus*.

Mesch, J. (2016). *Dataset. Tactile Swedish Sign Language Corpus*. Stockholm: Department of Linguistics, Stockholm University.

Nicodemus, B., Swabey, L., Leeson, L., Napier, J., Petitta, G., & Taylor, M. M. (2017). A cross-linguistic analysis of fingerspelling production by sign language interpreters. *Sign Language Studies* 17(2): 143–171.

Raanes, E. & Mesch, J. (2019). Dataset. Parallel Corpus of Tactile Norwegian Sign Language and Tactile Swedish Sign Language. Trondheim: Norwegian University of Science and Technology.

Schönstrom, K. & Mesch, J. (2017). *Dataset. The project 'From speech to sign – learning Swedish Sign Language as a second language'*. Stockholm: Department of Linguistics, Stockholm University.

Stokoe, W. (1960). Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *Studies in Linguistics: Occasional papers* 8.

Perniss, P., Pfau, R. & Steinbach, M. (eds). (2007). *Visible Variation: Cross-Linguistic Studies in Sign Language Structure*. Berlin: Mouton de Gruyter.

Vermeerbergen, M. & Nilsson, A. L. (2018). Introduction. In *A Bibliography of Sign Languages, 2008 – 2017*, A. Aarssen, R. Genis and E. van der Veken (eds). Leiden: Brill.

Roundtable discussion

Towards enriched corpus-based research designs for 21st century contrastive and translation studies – Roundtable discussion convened by Silvia Bernardini and Stella Neumann

Moritz Schaeffer[1], Haidee Kotze[2], Bert Le Bruyn[2], Marie-Aude Lefer[3], Silvia Bernardini[4], Stella Neumann[5]

[1]Mainz University, Germany, [2] Utrecht University, the Netherlands, [3]UCLouvain, Belgium, [4]University of Bologna, Italy, [5]RWTH Aachen University, Germany

The 1990s brought about a host of corpus-based studies of translation (CBTS), making it one of the most dynamic empirical approaches in translation studies. The main topic addressed by these studies has been the investigation of the features that characterise translation as a mode of communication. While obviously central to translation studies, this topic is also fundamental for contrastive linguistics, as use of translated data in this field relies on an awareness of its specificities with respect to non-translated language use.

These studies have provided a very substantial, precious body of data and hypotheses, yet their findings have also often turned out to be inconclusive and/or contradictory and/or difficult to explain.

We would argue that the main research questions that inspired these efforts have still not been answered satisfactorily, due, among other things, to design and methodological limitations in the first generation translation corpora. At the same time, the nature of translation has changed substantially in the last three decades, as a result of a multiplicity of factors, including ever less clear-cut boundaries between what counts as an 'original' and a 'translation', massively distributed collaborative processes, and human translation 'augmented' by machine translation in various ways. Recent next generation corpus compilation efforts have taken steps to resolve these shortcomings, but much more remains to be done in order to grapple with the complexity of the topic.

This is therefore the right time to take stock of corpus-based research designs in translation studies and review possible ways ahead for CBTS towards fully exploiting the potential of the corpus approach, towards richer explanations in translation and contrastive studies.

To this end, the INTERACT working group on enriched research designs organises a round table that will address the following and further related questions:

- innovative designs of corpora for translation and contrastive studies
- beyond the comparable/parallel dichotomy
- challenges with multi-parallel corpora
- sampling bias versus coverage of translation-related types of texts
- ecological validity versus elicitation
- cross-linguistic comparability of data sets
- rich metadata
- capturing behaviour/cognition through corpora

Arabic sentence patterns in interpreted, translated and original speeches: A corpus-based approach

Aladdin Al Zahran

Sohar University, Sultanate of Oman

KEYWORDS: contrastive interpreting studies, structural asymmetry, corpus-based investigation, simultaneous interpreting, Arabic

This paper reports on a corpus-based implementation of contrastive interpreting studies. In simultaneous interpreting (SI) from a subject-verb-object (SVO) language such as English, Arabic simultaneous interpreters are confronted with the dilemma of choosing between the default, dominant, and unmarked verb-initial or VS(O) structure on the one hand, or a derived and *marked* subject-initial or SV(O) structure on the other hand. The choice of the default VS(O) structure may involve a risk of cognitive overload and potential information loss or failure due to the need to lag behind the original speaker. This need is even more pronounced if this structural asymmetry coincides with other extreme factors such as high source language (SL) input rate, structural complexity (e.g., in the subject position) and/or information density. On the other hand, the choice of any of the SV(O) structures available in Arabic would help the simultaneous interpreter in closely following the structure of SL discourse without the need to lag far behind the original speaker despite the shift of focus and emphasis brought about by the marked SV(O) structure.

This research builds on a body of literature investigating language-pair-specific factors, language-pair specificity and strategies and tactics employed by professional interpreters to cope with structural asymmetry.

The main objective of this paper is to determine the structure used more frequently in simultaneously interpreted, translated and original speeches.

To achieve the research objective, a multi-layered analysis is conducted on a unidirectional English-to-Arabic parallel multimodal corpus with a combination of transcribed mediated (simultaneously interpreted) and non-mediated (original) spoken material, as well as written translations of mediated speeches. The transcribed material comprises three sub-corpora as follows: **Sub-corpus (A)**, which comprises transcriptions of seven source political English speeches; **Sub-corpus (B)**, which consists of transcriptions of multiple Arabic SI performances of the source English speeches by 21 interpreters interpreting into their mother tongue; and **Sub-corpus (E)**, which consists of transcriptions of original Arabic speeches used as a reference corpus. **Sub-corpus (C)**, in turn, contains multiple written translations of the source English speeches in Sub-corpus (A) performed by eight professional interpreters.

The original English and Arabic speeches belong to the specific genre of political speeches. Along with the SIs and translations of the English speeches, they form a highly specialised type of language. The more highly specialized the language to be sampled in the corpus, the fewer will be the problems in defining the texts to be sampled. To satisfy the representativeness requirement, the selection criteria stipulate the following:

1. Any political English speech delivered in a formal or conference-like setting, whether at the national and/or international level, and for which a professional Arabic SI is available will be a candidate for inclusion in the corpus.
2. Moreover, any political Arabic speech delivered by a native Arab official using Modern Standard Arabic, the default and expected Arabic variety in conference interpreting and media settings, in a

formal or conference-like setting, whether at the national and/or international level, will be a candidate for inclusion in the corpus.

3. Additionally, the original Arabic speeches must have been delivered from 1970 onwards. The 1970 landmark slightly exceeds 50 years and has been chosen to achieve contemporaneousness and parallelism with the English speeches.

Currently, the corpus comprises **17:19:11 hours** of English speeches, their Arabic SI performances and translations, as well as original Arabic speeches with a total size of **107,625** tokens.

The analysis of the structures used mostly in written translations and original Arabic speeches is expected to help interpret the Arabic SI analysis results and relate them to the SI process. In particular, due to the fact that Arabic SV(O) structures would be easier to process than VS(O) ones, we hypothesize that the Arabic simultaneous interpreters would opt for the former rather than the latter despite the risk of shift in emphasis involved in such a structural choice. We also hypothesize that the opposite will be true in the case of written translations and original Arabic speeches because, not being under the type of cognitive constraints characteristic of SI, the professional translators and Arab speakers will opt for the default and unmarked VS(O) structure.

The preliminary results of our multi-layered corpus-based analysis confirm our hypotheses as they indicate that the dominant structure in the Arabic SI performances analysed was the SV(O) structure while in translations and original Arabic speeches, the VS(O) structure was the dominant one. The results of the study indicated that the choice of structure was not based on a single aspect or modality (mediated vs. non-mediated, or spoken vs. written), but a combination of two aspects. Thus, if our sub-corpora were placed on a continuum with the SVO structure at one extreme and the VSO structure at the other, we will notice that the more mediated and spoken the discourse the closer it will be to the SVO structure. Conversely, the more mediated and written or non-mediate and spoken the discourse, the closer it will be to the VSO structure.

These preliminary results may indicate that, when confronted with the asymmetrical SV(O) structure in English>Arabic SI, the simultaneous interpreters resort to «form-based processing» as opposed to «sense-based processing» to avoid restructuring. This result is clearly evidenced by the interpreters' tendency to follow the Arabic structure that is similar to the English SVO structure as a coping tactic with structural asymmetry. These results may thus provide support to the hypothesis of «language-pair specificity» since SI between languages with similar or flexible structures does require restructuring.

References

Abdul-Raof, H. (1998). Subject, theme and agent in Modern Standard Arabic. Routledge.

Al-Rubai'i, A. M. (2004). The effect of word order differences on English-into-Arabic simultaneous interpreters' performance. *Babel, 50*(3), 246-266.

Al Zahran, A. (2021). Structural challenges in English>Arabic simultaneous interpreting. *Translation & Interpreting, 13*(1), pp. 51-70. doi: 10.12807/ti.113201.2021.a04

Al Zahran, A., & Jamoussi, R. (2022). A corpus-based investigation of VSO-SVO usage in simultaneous interpreting. *Hikma 21*(2), 231-255. https://doi.org/10.21071/hikma.v21i2.14281

Alonso Bacigalupe, L. (2010). Information processing during simultaneous interpretation: A three-tier approach. *Perspectives: Studies in Translatology, 18*(1), 39- 58. https://doi.org/10.1080/09076760903464278

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1-16.

Bartłomiejczyk, M., Gumul, E., & Koržinek, D. (2022). EP-Poland: Building A bilingual parallel corpus for interpreting research. *GEMA Online Journal of Language Studies, 22*(1).

Bendazzoli, C., & Sandrelli, A. (2009). Corpus-based interpreting studies: Early work and future prospects. *Revista Tradumàtica*, 7.

Castagnoli, S. (2011). Exploring variation and regularities in translation with multiple translation corpora. *Rassegna Italiana di Linguistica Applicata*, *43*(1), 311-332. doi:10.1400/190460.

Collard, C., Przybyl, H., & Defrancq, B. (2018). Interpreting into an SOV language: Memory and the position of the verb. A corpus-based comparative study of interpreted and non-mediated speech. *Meta, 63*(3), 695-716. https://doi.org/10.7202/1060169ar

Gile, D. (2011). Errors, omissions and infelicities in broadcast interpreting: Preliminary findings from a case study. En C. Alvstad, A. Hild, & E. Tiselius (Edits.), *Methods and strategies of process research: Integrative approaches in translation studies* (pp. 201–218). John Benjamins.

Goldman-Eisler, F. (1972). Segmentation of input in simultaneous translation. *Journal of Psycholinguistic Research, 1*(2), 127–140.

Dam, H. V. (2001). On the option between form-based and meaning-based interpreting: The effect of source text difficulty on lexical target text form in simultaneous interpreting. *The Interpreters' Newsletter, 11*, 27–55.

He, H., Boyd-Graber, J., & Daumé III, H. (2016). Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. *En HLT-NAACL* (pp. 971–976).

Isham, W. P. (1994). Memory for sentence form after simultaneous interpretation: Evidence both for and against deverbalization. En S. Lambert, & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 191–211). John Benjamins.

Kirchhoff, H. (1976/2002). Simultaneous interpreting: Interdependence of variables in the interpreting process, interpreting models and interpreting strategies. In F. Pöchhacker, & M. Shlesinger (Eds.), *The interpreting studies reader* (pp. 111–119). Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Moser, B. (1978). Simultaneous interpretation: A hypothetical model and its practical application. In D. Gerver, & H. W. Sinaiko (Eds.), *Language interpretation and communication* (pp. 353–368). Plenum Press.

Setton, R. (1999). Simultaneous interpretation: A cognitive-pragmatic approach. John Benjamins.

Setton, R., & Motta, M. (2007). Syntacrobatics: Quality and reformulation in simultaneous-with-text. *Interpreting, 9*(2), 199–230.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (Vol. 6). J. Benjamins.

Van Besien, F. (1999). Anticipation in simultaneous interpretation. *Meta, 44*(2), 250–259.

Wang, Binhua, & Gu, Yukui. (2016). An evidence-based exploration into the effect of language-pair specificity in English-Chinese simultaneous interpreting. *Asia Pacific Translation and Intercultural Studies*. https://doi.org/10.1080/23306343.2016.1182238

Wang, B., & Zou, B. (2018). Exploring language specificity as a variable in Chinese-English interpreting. A corpus-based investigation. In M. Russo, C. Bendazzoli, & B. Defrancq (Eds.), *Making way in corpus-based interpreting studies* (pp. 65-82). Springer.

Dative alternation in Italian-to-Dutch translation: a corpus-based study on the effect of structural priming

Francesca Antonioli, Gert De Sutter, Jelena Vranjes, Joke Daems

Ghent University, Belgium

KEYWORDS: structural priming, Italian-to-Dutch translation, dative construction, DOC, POC

Structural priming is the tendency to repeat the same syntactic structure of a recently comprehended or produced sentence. The notion of structural priming was introduced in 1986 by Bock; since then, it has been extensively investigated in both intra- and cross-linguistic studies with experimental designs or observational approaches relying on corpus-based data (for an overview, see Pickering & Ferreira, 2008 and Gries & Kootstra, 2017). Recently, corpus-based research on cross-linguistic structural priming in bilinguals has also been applied to empirical translation studies, with the aim of investigating whether the translator is primed to repeat syntactic structures of the ST in the TT. The studies by Bangalore et al. (2016), De Sutter et al. (2021) and De Sutter et al. (2023), e.g., point in this direction.

With this paper, we aim to contribute precisely to the line of research on the effects of cross-linguistic structural priming on translation by empirically investigating the dative alternation in untranslated Dutch texts and in Italian-to-Dutch translated texts, i.e. the variation between the double object construction (DOC) and the prepositional object construction (POC). More specifically, we want to find out whether and to what extent the Italian dative construction of the ST influences the translator's free choice between DOC and POC in the writing of the Dutch TT, while controlling for other explanatory variables that have been shown influential in previous literature.

To answer this research question, we adopted a corpus-based approach and we relied on *InterCorp*, a multilingual parallel corpus available on the Czech platform KonText of Charles University (Boková et al., 2021). We considered InterCorp to be the most suitable corpus for the language pair at hand (Italian and Dutch): it contains metadata about ST and TT as well as genre; it is POS-tagged and syntactically parsed, and the results are displayed with different types of metatextual information, including the author's name and the publication year of the text. We limited the number of texts to be analysed, selecting only the genres *fiction* (n= 6,506,929), *journalism commentaries* (n= 3,079,659) and *journalism news* (n= 942,023). Of all the genres available in InterCorp, these allow us a comparison with Van Beveren's doctoral thesis (2021), devoted to dative alternation in untranslated Dutch and Dutch translated from French or English. We used two different CQL queries to obtain the Italian-to-Dutch data: firstly, we queried the subcorpus for all indirect objects that are syntactically dependent on the Dutch verbs that allow dative alternation (we selected from the list in Colleman, 2006 the 20 verbs with the highest total frequency in the DO and the PO construction together); secondly, we queried all nouns preceded by the preposition *aan* which are dependent on the same verbs. We automatically extracted all results, which, prior to manual validation, consisted of 615 instances of DOC and 319 instances of *aan*-phrases. To compare translated and untranslated texts, we used the same queries to extract data from a subcorpus consisting of the same text-types, this time however with Dutch as source language (representing untranslated Dutch). Before manual validation, untranslated Dutch data consisted of 130 instances of DOC and 92 instances of *aan*-phrases. All instances of this first dataset were manually checked, and irrelevant attestations were removed. Relevant instances contained a DOC or POC replaceable by its alternative without leading to ungrammaticality.

The final dataset included 719 Italian-to-Dutch instances (565 DOC and 154 POC) and 169 untranslated Dutch instances (114 DOC and 55 POC). Each one of these instances was manually annotated for the response variable *Dutch dative construction* (DOC vs. POC), the *source language* (non-translated vs. Italian), and Italian-to-Dutch instances were also annotated based on the type of dative construction found in the Italian source text (*Italian dative construction*). Since the existence of dative alternation in Italian is a debated issue, the variable *Italian dative construction* needs some more specific clarification. For the purposes of this study, nominal datives and pronominal datives with tonic personal pronouns in Italian

source texts were annotated as POC, because they are obligatorily preceded by the preposition *a* (Folli et al., 2006). In contrast, pronominal datives with atonic (or clitic) personal pronouns were annotated as DOC, because they are never preceded by any preposition. To consider possible counter explanations of the dative construction other than structural priming, additional annotations were added for some predictor variables that, according to the literature, constrain dative alternation (i.e., *pronominality*, *animacy*, *difference in length between patient and recipient* and *genre).*

We ran two logistic regression analyses, using the statistical software package R (R Core Team, 2022). The first analysis aimed at finding out to what extent the predictors that are not related to structural priming, i.e. *pronominality, animacy, length difference* and *genre*, have a different effect in translated vs. untranslated Dutch. Therefore, we included the mentioned predictors as main effects as well as in two-way interaction effects with *source language* in the regression model. The second analysis, then, focussed on translated Dutch only, with the aim of finding out to what extent the structural priming variable influences the response variable vis-à-vis the other predictor variables. The results of the first analysis show that there are no significant two-way interactions, but there are three significant main effects, i.e. *pronominality, animacy* and *length difference between patient and recipient*. More specifically, when the dative is pronominal, animate, and shorter than the patient, it has a much higher probability of being rendered with a DOC; whereas, when the dative is nominal, inanimate and longer than the patient, it is more likely to be used in a POC. Given the lack of any interaction effect, we conclude that the predictors affect the choice between the dative constructions in an identical way in translated and untranslated Dutch. The second analysis shows that in translated Dutch, the dative construction is not only influenced by *animacy, pronominality* and *length difference*, but also shows a strong priming effect: a DOC in the Italian source text increases the likelihood of having a DOC in the Dutch target text, and vice versa for the POC. This conclusion mirrors the findings of previous corpus-based studies on cross-linguistic priming in translation, such as De Sutter et al. (2021) and De Sutter et al. (2023).

In our presentation, we will illustrate the results of both regression analyses in more detail, we will further explore the impact of structural priming in a process study (using keylogging and eyetracking) of Italian-to-Dutch translations by student translators, and we will elaborate on the theoretical status of structural priming in an empirical-based theory of translation.

References

Bangalore, S., Behrens, B., Carl, M., Ghankot, M., Heilmann, A., Nitzke, J., Schaeffer, M., & Sturm, A. (2016). Syntactic variance and priming effects in translation. *New directions in empirical translation process research: Exploring the CRITT TPR-DB*, 211-238.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, *18*(3), 355-387.

Boková, E.; Hrnčířová, Z.; Vavřín, M.; Rosen, A.: *Korpus InterCorp – nizozemština, verze 13ud z 22. 12. 2021*. Ústav Českého národního korpusu FF UK, Praha 2021. http://www.korpus.cz

Colleman, T. (2006). De Nederlandse datiefalternantie: een constructioneel en corpusgebaseerd onderzoek/door Timothy Colleman (Doctoral dissertation, Ghent University).

De Sutter, G., Colleman, T., & Ghyselen, A. S. (2021). Intra-and Inter-textual Syntactic Priming in Original and Translated English. *Applications of Cognitive Linguistics*, 410.

De Sutter, G., Lefer, M.-A., & Vanroy, B. (2023). Is linguistic decision-making in student translation constrained by the same cognitive factors as in professional translation? Evidence from subject placement in French-to-Dutch news translation. *International Journal of Learner Corpus Research.*

Folli, R., Harley, H., Doetjes, J., & Gonzáez, P. (2006). Benefactives aren't Goals in Italian. *Amsterdam studies in the theory and history of linguistic science series 4*, *278*, 121.

Gries, S. T., & Kootstra, G. J. (2017). Structural priming within and across languages: A corpus-based perspective. *Bilingualism: Language and Cognition*, *20*(2), 235-250.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: a critical review. *Psychological bulletin*, *134*(3), 427.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Van Beveren, A. (2021) De verklaringen voor toegenomen explicietheid in vertalingen: een multifactorieel corpusonderzoek naar de Nederlandse om-alternantie en datiefalternantie (Doctoral dissertation, Ghent University).

Internationalizing a mega-university: a corpus-based study on the translation of different text-types at the University of Bologna

Gaia Aragrande and Ettore Galletti

University of Bologna, Italy

KEYWORDS: Institutional translation, Internationalizing universities, Translation strategies, Style and register, Corpus analysis

This paper presents our personal experiences as research fellows working at the International Relations Strategy Support Unit (IRSS) at the University of Bologna. We will centre our analysis on the translational activities carried out over a period of 12 months between 2020 and 2021 and that involved three different text types: news articles (press releases) and communications to the university community by members of the university governance; journalistic texts (research outreach articles) published in Italian and English on the University magazine (UniboMagazine); social media posts issued through the University's institutional Facebook and Instagram accounts. We carried out these translational activities using Trados Studio along with MultiTerm, allowing us to record and keep our translation memories in order and, more importantly, to build two separate parallel corpora (UNIBO_Comm, UNIBO_Social) that will serve as basis of our analysis.

Our observations will consider the translation products (our translations), the external factors influencing our performance as well as the contexts of IRSS and the central governance of the University of Bologna. On the one hand, we will use the corpora above to investigate translation strategies, mistranslations and to define our general attitude to the source texts while also attempting to answer the following research questions: are there some text type-specific translation strategies that we employed? Does our text type-specific attitude reflect what has been explored by existing literature in the field of institutional translation?

On the other hand, we will try to intersect issues generally pertaining to institutional translation where "process, competence and product are inextricably intertwined" (Ramos, 2015, pp.23) with the specificities characterizing the University of Bologna. The latter is generally defined as a mega-university due to the number of its students. However, its size does not uniquely refer to the student community, but it extends to the high number of employees and stakeholders who are daily involved in the university activities as a higher education institution and as a prominent player at the local and international level. In this, the University of Bologna is always addressing multiple audiences in terms of competence, roles, access and interests and it does so, at present, without a clear set of internal guidelines and best practices for communicating to its own international public.

The two corpora under study are composed as follows:

- UNIBO_Comm: 36,890 words
- UNIBO_Social: 42,267 words

These two corpora are comparable in terms of size, but they do present some differences that go beyond word count and have to do with the text types they include and their distribution over time. Indeed, as shown in Figure 1, the UNIBO_Social corpus is larger in size but also more consistent in terms of both text types and diachronic distribution. Indeed, this corpus only comprises social media posts on two different platforms (Facebook and Instagram) that present different communicative needs thus justifying the different plotting over time and according to social media platform (Figure 2 and 3).

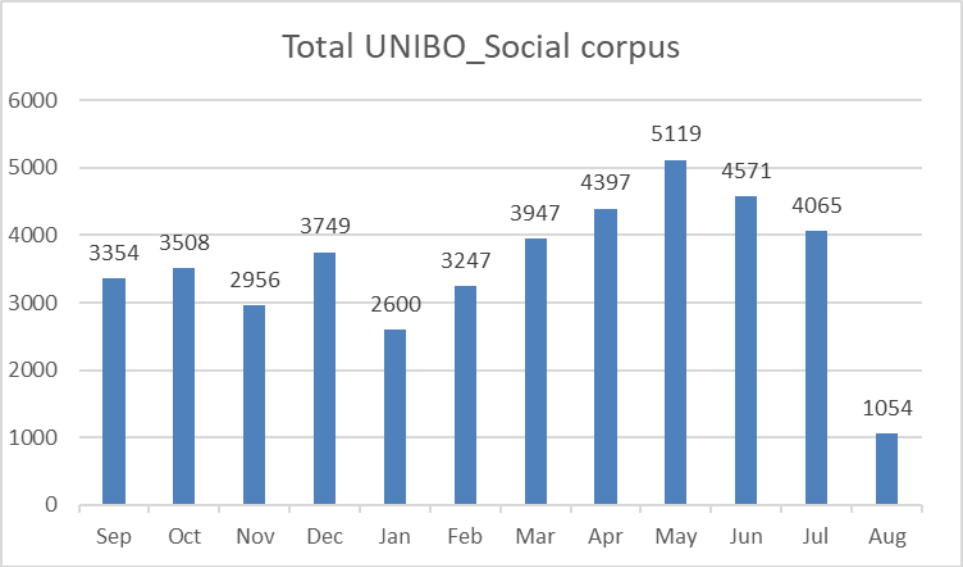


Figure 1. Monthly distribution of translated social media posts in the whole UNIBO_Social corpus

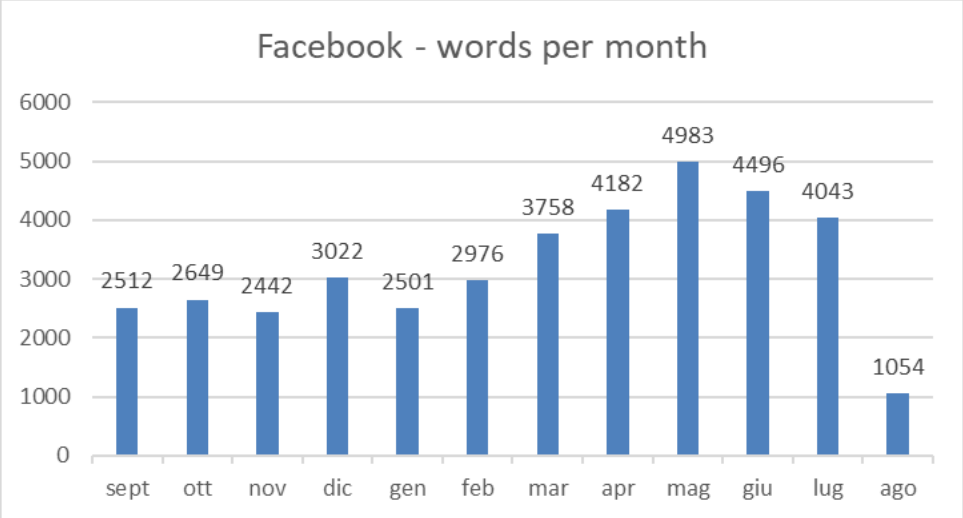


Figure 2. Monthly distribution of translated Facebook posts in the UNIBO_Social corpus

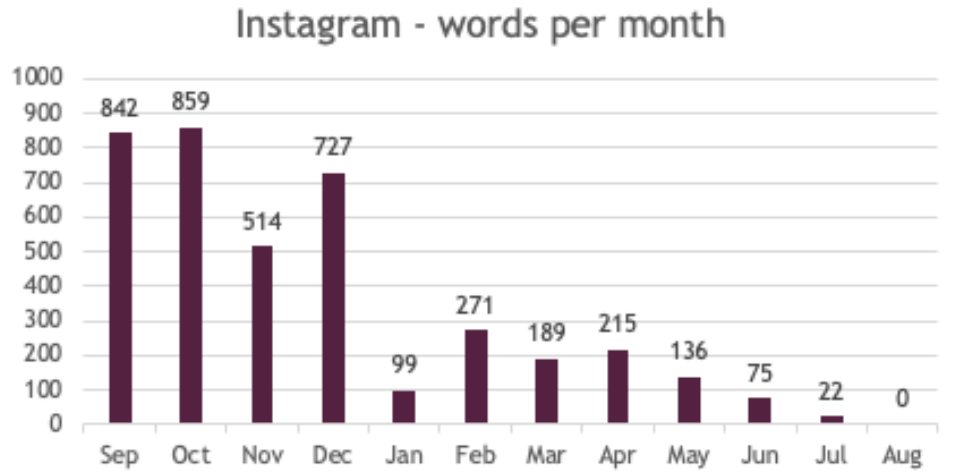


Figure 3. Monthly distribution of translated Instagram posts in the UNIBO_Social corpus

The situation is quite different in the UNIBO_Comm corpus which comprehends three different text types:

1) Press releases (Comunicati) published either on the UniboMagazine or on the University website in the "Notice" section;
2) Research outreach reports (Ricerca) published on the UniboMagazine;
3) Communications to the community (Infoateneo) sent via email in double language or published on the University website in the "Notice" section.

The plotting of texts/words over time and across text types varies a lot in this corpus. By looking at Figure 4, we can appreciate how texts in the UNIBO_Comm corpus present a striking majority of research outreach reports (Ricerca) in the first six months considered in the corpus (Sept 2020 – Feb 2021), while, during the following six months (Mar 2022 – Aug 2022) we witness to a general decrease (except a few exceptions) in the number of translated words of every text type and especially of research reports.

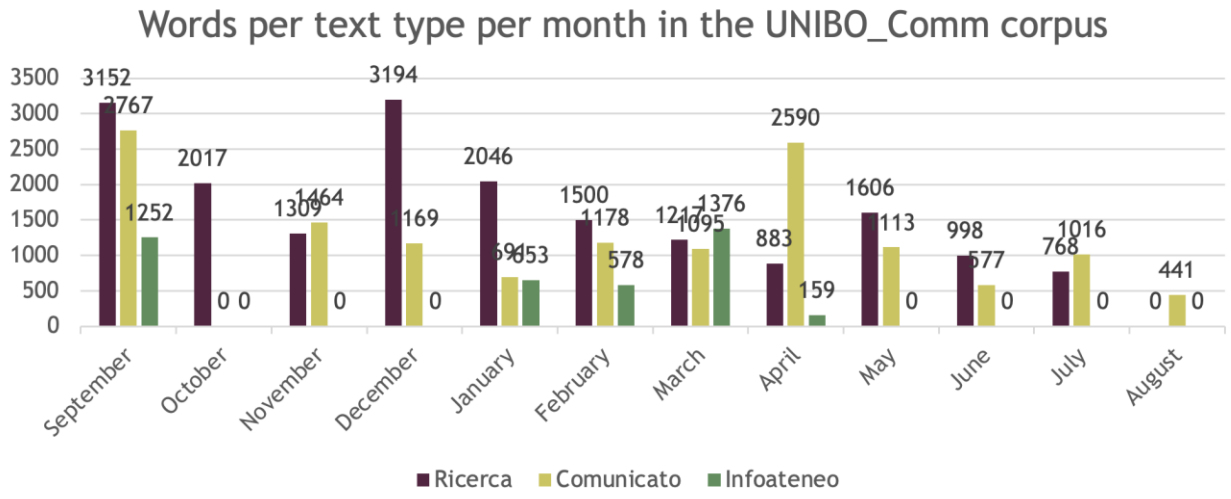


Figure 4. Monthly distribution of words per text type in the UNIBO_Comm corpus

Our decision to keep all the texts of the UNIBO_Comm in one single corpus was mainly practical: these text types are somehow comparable and deal with what can be understood as "news" from the University of Bologna. Moreover, to preserve the integrity of the case study, we thought that having two corpora ensured a greater inter-corpora comparability, despite such an uneven distribution both text type-wise and time-wise. However, we opted for a tagging system that allows us, if needed, to divide these texts in sub-corpora, each for every text type or social media platform.

Besides trying to answer the abovementioned research questions with traditional corpus investigation methods (parallel concordance lines, keywords extraction in the source and target language and collocational analysis), we also attempt to make hypotheses as to why our corpora present such uneven distributions over time and across text types. Both corpora show some differences in these regards and the reasons are to be found outside the corpora and the translation activity and inside the contextual framework of the University of Bologna.

As far as the abovementioned research questions are concerned, some primary results confirm our initial hypothesis of having very different attitudes towards style and register according to the communication typology, but we witnessed to a hybridization of registers when it comes, for example, to social media posts reporting or reposting research outreach news. In these cases, translators actively intervened in the lead of the research outreach report to make it less formal and catchier when it was posted on a social media platform. A further preliminary result concerns differences found in the translation of press

releases. Indeed, we observed how if the press release reported or was based on a communication previously sent to the entire community, translators tended to replicate the style and register of such communication. There might be many reasons justifying this more conservative attitude on the part of the translators, our hypothesis has to do with some kind of authorial stance of the message itself. Indeed, these messages to the community were always signed either by the Rector or one of the Vice-Rectors – therefore a more ST-oriented approach was adopted. The press releases containing or reporting information taken from such messages were perhaps influenced by this authoritative presence.

References

Mikhailov, M., & Cooper, R. (2016). Corpus linguistics for translation and contrastive studies: A guide for research. London and New York: Routledge.

Prieto Ramos, F. (2015). Quality Assurance in Legal Translation: Evaluating Process, Competence and Product in the Pursuit of Adequacy. *International Journal for the Semiotics of Law, 28*(1), 11–30.

Prieto Ramos, F. (ed. 2018). Institutional Translation for International Governance: Enhancing Quality in Multilingual Legal Communication. London: Bloomsbury.

Corpus approaches to news translation: can we do better than comparable?

Silvia Bernardini, Adriano Ferraresi, Federico Garcea, Natalia Rodriguez Blanco

University of Bologna, Italy

KEYWORDS: news translation, comparable corpora, parallel corpora, machine translation, sentence embeddings

This contribution addresses the challenging issue of building corpus resources for the study of news translation. It focuses specifically on global news agencies as a discourse production setting in which translation plays a fundamental, though often unacknowledged, role.

We propose a method for identifying translated segments within multilingual news dispatches that relies on machine translation and semantic similarity scores to determine what text units carry traces of translation within trilingual sets of dispatches issued by the same news agency (Agence France-Press, AFP) in Spanish (ES), French (FR) and English (EN).

The implications of the problem at hand impact on the blurring separations between contrastive vs. translation studies (TS) and parallel vs. comparable corpus architectures, on fundamental theoretical notions related to equivalence, authorship and translation units, and ultimately on the status of translation as a socially relevant discourse production practice.

1. Previous work

1.1. News translation and global news agencies

Within TS, the subfield of ‘journalistic translation’ (Valdeón, 2015) or 'news translation' (Holland, 2013), has grown in the last two decades. Yet the most notable trait of this field of study remains its flimsiness: translation in news settings is largely invisible, text producers do not consider themselves as translators, and their texts are characterised ‘by a high degree of transformation and rewriting’ (Bielsa, 2010: 48). Indeed, the texts they produce might be considered ‘far’ from ‘an accurate translation that is true to the original’ (Hernandez Guerrero, 2022: 232), being characterised by the coexistence of contradictory practices such as radical rewriting and literal translation (Davier, 2021).

Studies about ‘translational phenomena in the news’ (Davier, 2022) thus pose theoretical and methodological challenges of more general import. Basic steps in the study of ‘prototypical’ translation, such as tracing down source and target texts (Caimotto & Gaspari, 2018) and defining equivalence, translation units and even authorship (Gambier, 2022; Bielsa & Bassnet, 2009) are far from straightforward in this field.

Against the backdrop of news translation, this study focuses on the multilingual output of the global news agency AFP. Considered as one of the most influential agencies, AFP publishes news in French, English, Spanish, Arabic, German, and Portuguese (Bielsa & Bassnett, 2009). Its headquarters in Uruguay centralise the regional news coverage and translation services dealing with Spanish (Rodriguez-Blanco, forthcoming).

1.2. Bridging the gap between comparable and parallel corpora

Corpus studies of news translation typically rely on multilingual comparable corpora, given the difficulty of identifying source-target pairs in this domain (Davier & van Doorslaer, 2018), and of singling out sub-textual translation units within them (Gambier, 2022). While the first issue can be partly addressed using paratextual and ethnographic approaches, identifying translation units remains problematic, hindering the development of parallel corpora for this heavily mediated translation type. This problem has been addressed in two main ways.

Computational techniques have been employed to mine parallel sentences from comparable corpora and noisy parallel corpora to increase the amount of training data for machine translation systems (Barrón-Cedeño et al., 2015, Gete & Etchegoyhen, 2022). Less computationally intense tasks have instead led to the creation of hybrid *comparallel* corpora. The term has been used to refer to collections of texts aligned at the document level, where the actual correspondence of translation units at the sub-textual (paragraph or sentence) level, if any, is left for the researcher to work out (Bernardini et al., 2010, Gaspari, 2015).

The question that we address with this contribution is whether we can do better than comparable corpora in news translation research, given that, in the words of Gaspari and Caimotto (2018, p. 216) "adopting straightforward conventional parallel corpus methodologies seems hardly feasible in most news translation scenarios".

2. Method

We selected one dispatch in three versions (ES>FR>EN) that we identified based on paratextual and ethnographic information as likely to have been, at least to some extent, translated. The dispatches are around 650-word long and deal with the 2020 presidential elections in Bolivia.

To obtain benchmark data for the automated evaluation, and assess the feasibility of the task, we carried out a manual evaluation, categorizing each sentence as either 'translated' or 'not translated'. Three authors carried out the evaluation independently. Interrater agreement as measured by Krippendorff's α varied for the three language pairs (ES>FR: 0.714; ES>EN: 0.648; FR>EN: 0.605), but in all cases can be considered as substantial (cf. Arstein & Poesio, 2008, p. 576).

An English translation of the FR and ES dispatches was obtained using the MT system ModernMT and similarity scores between each pair of sentences in each pair of documents were computed. We used HuggingFace's Sentence Transformer library (Reimers & Gurevych, 2019) to compute embeddings for each sentence pair and entered them in similarity score matrices displaying them as heat maps.

4. Preliminary results

As a preliminary evaluation we matched the automatic similarity scores assigned to each sentence against the manual evaluation. Sentences receiving a score above 0.7 tend to be unanimously recognized as (semi-)parallel, while a higher degree of uncertainty on the part of the evaluators is associated with lower scores (Fig. 1).

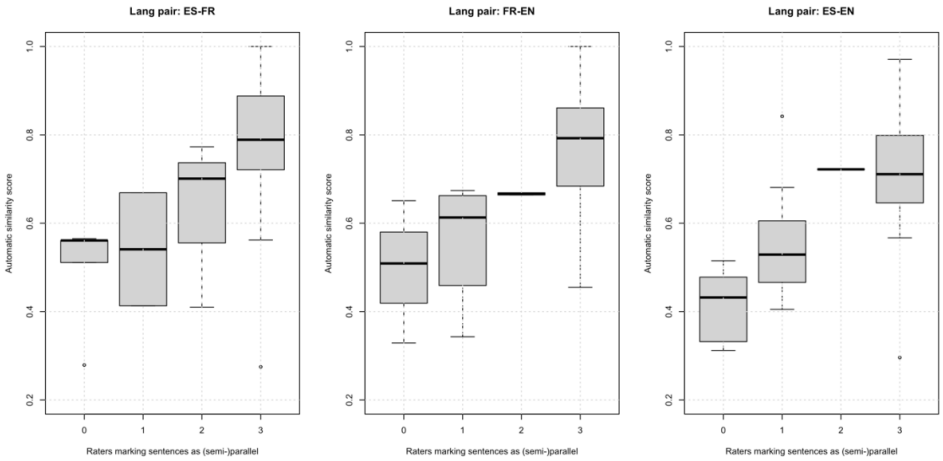


Figure 1. Comparison between Automatic Scores and Manual Evaluations

5. Discussion and conclusion

In this contribution we address the issue of identifying sub-textual translation units within trilingual news dispatches identified through paratextual and ethnographic cues as being translation candidates. Our method relies on the computation of sentence embeddings and similarity scores between a machine translated version of a dispatch into a given language, and its corresponding published version in the same language.

We show that these similarity scores are in line with those obtained through a non-trivial human evaluation task. Scores >0.8 would seem to point to alignable units and scores around 0.7 to partial matches.

After refining this method and comparing it with others based on cross-language similarity, we aim to set up a prototype *comparallel* corpus that can be accessed both from the comparable and parallel perspectives. While this is only a first step, resources of this kind would allow us to go beyond the artificial boundaries that have so far contributed to hiding the role of translation as a discourse creation device in numerous multilingual settings.

References

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Bielsa, E. (Ed.). (2021). The Routledge Handbook of Translation and Media. Routledge.

Bielsa, E. (2010). Translating News: A comparison of Practices in News Agencies. In R. Valdeón, R. (Ed.), *Translating information* (31–49). Universidad de Oviedo.

Bielsa, E., & Bassnett, S. (2009). *Translation in global news*. Routledge.

Caimotto, C. & Gaspari, F. (2018). Corpus-based study of news translation: Challenges and possibilities. *Across Languages and Cultures*, *19*(2), 205–220.

Davier L. (2021). Translation in the news agencies. In E. Bielsa (Ed.), *The Routledge Handbook of Translation and Media* (183–198). Routledge.

Davier, L. (2022). Translational phenomena in the news: Indirect translation as the rule. *Target. International Journal of Translation Studies*, *34*(3), 395–418.

Filmer, D. (2014). Journalators? An ethnographic study of British journalists who translate. In *Cultus. The Intercultural Journal of Mediation and Communication 7*, 135–158.

Gambier, Y. (2022) Revisiting certain concepts of translation studies through the study of media practices. In E. Bielsa (Ed.), *The Routledge Handbook of Translation and Media* (91–107). Routledge.

Gambier, Y., & van Doorslaer, L. (2016). Disciplinary dialogues with translation studies: The background chapter. In Y. Gambier & L. van Doorslaer (Eds.), *Benjamins Translation Library* (Vol. 126, pp. 1–22). Benjamins.

Hernández Guerrero, M. J. (2022). The translation of multimedia news stories: Rewriting the digital narrative. *Journalism*, *23*(7), 1488–1508.

Holland, R. (2013). News translation. In C. Millán & F. Bartrina (Eds.), *The Routledge Handbook of Translation Studies* (pp. 332–346). Routledge.

Hursti, K. (2001): An insider's view on transformation and transfer in international news communication: An English-Finnish perspective. *The electronic journal of the Department of English at the University of Helsinki*, *1*, 1–8.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Rodriguez-Blanco, N. (forthcoming). Plurilingual Perspectives, Pluricultural Contexts. In *Exchanges Interdisciplinary Research Journal*. University of Warwick.

Stetting, K. (1989). Transediting. A new term for coping with the grey area between editing and translating. In G. Caie, et al. (Eds.), *Proceedings from the Fourth Nordic Conference for English Studies* (pp. 371–382). University of Copenhagen.

Valdeón, R. (2015). Fifteen years of journalistic translation research and more. *Perspectives 23*(4), 634–662.

Errors in student translation and post-editing: same or different?

Romane Bodart & Marie-Aude Lefer

UCLouvain, Belgium

KEYWORDS: learner translation corpora, machine translation post-editing, computer-aided error annotation, translation quality, translator education

While the bulk of corpus-based translation research to date has examined professional translations, attention is now increasingly being directed to other forms of translation such as learner (or student) translations, i.e. translations produced by translation students or foreign language learners. The compilation of learner translation corpora is not new. Pioneering projects include PELCRA (Uzar, 2002) and the Student Translation Archive (Bowker & Bennison, 2003). These early initiatives were soon followed by similar corpus projects, such as MeLLANGE (Castagnoli et al., 2011) and RusLTC (Kutuzov & Kunilovskaya, 2014), with learner translation corpus research steadily gaining momentum. Today, it is fair to say that learner translation corpus research has a home of its own in the corpus landscape. In a recent survey, Granger & Lefer (2023) show that quality evaluation and computer-aided error analysis hold center stage in learner translation corpus studies. Error-annotated corpora can be used for a wide range of applied and theoretical purposes (e.g. materials design, data-driven learning, studies into translation competence acquisition). A new development in learner translation research, which goes hand in hand with the technological turn we are witnessing, is the compilation and error-annotation of student post-editing corpora, i.e. parallel corpora containing machine translations post-edited by students (see e.g. Kübler et al., 2022).

In this talk, we aim to provide comparative insights into the translation quality of two types of student production, namely translations from scratch and machine-translation post-edited texts (see e.g. Garcia, 2010 and Daems et al., 2017 for experimental research on the topic). Our objective is two-fold: (1) assess the overall quality of translations and post-edited texts produced by the same students and (2) compare the error types most commonly found in the two modes of production. The systematic comparison of translation and post-editing products is a much-needed endeavor in translator education and research into translation and post-editing competence acquisition. On the one hand, translation trainers are often left wondering when to introduce machine translation post-editing (PE) in translation curricula. A widely-held assumption is that students need to be properly trained in translation before PE can be practiced in class. On the other hand, it is often claimed that translation competences such as bilingual and extralinguistic competences support core post-editing competences (e.g. Nitzke et al., 2019), but little is known about the exact role they play in post-editing.

The present study is based on English-to-French corpus data drawn from the Multilingual Student Translation corpus (MUST; Granger & Lefer, 2020) and a sister project devoted to student post-editing, called postedit.me (PEM; Lefer et al., 2023). More precisely, it relies on MUST and PEM subcorpora containing respectively 56 translations (60,000+ tokens) and 56 post-edited texts (34,000+ tokens) produced by 23 second-year master's students trained in specialized translation and post-editing. The source texts in the two subcorpora are comparable and pertain to the legal and financial domains. The two subcorpora were error-annotated relying on two standardized taxonomies. The first, used to annotate student translations, is the Translation-oriented Annotation System (TAS) developed collaboratively within the MUST network (Granger & Lefer, 2021). TAS contains eight error categories: Content, Culture, Translation Brief, Grammar & Syntax, Lexis & Terminology, Discourse & Pragmatics, Mechanics, and Register & Style. The second, used to annotate student post-edited texts, is the Machine Translation Post-Editing Annotation System (MTPEAS; Lefer et al., 2022). The MTPEAS taxonomy contains seven categories. The first three, Value-adding edit (NE-PLUS), Unnecessary edit (NE-SUPF) and Error-introducing edit (NE-INTR), are used to tag segments in the PE that correspond to what were initially error-free MT segments (NE = 'no error' in the MT). The other four, Missing edit (E-MISS),

Successful edit (E-OKAY), Incomplete edit (E-INCO), and Unsuccessful edit (E-FAIL), are used to tag PE segments that correspond to erroneous MT segments (E = 'error' in the MT). The NE-INTR, E-MISS, E-INCO and E-FAIL categories are complemented with TAS tags to specify the nature of the erroneous PE segments. The datasets used in the present study contain 4,000+ MTPEAS annotations in PEM (comprising 1,200+ tags complemented with TAS annotations) and 1,200+ TAS error annotations in MUST. The annotations have been performed by one rater. Some of the texts have been annotated by a second rater to fine-tune the annotation process, but we have not computed inter-rater reliability scores at this stage.

In the study, we operationalize quality in translation and post-editing as the average number of errors per 1,000 tokens. Alongside translation method (translation vs post-editing), we examine the impact of three student and task variables encoded as MUST and PEM metadata, namely translation experience (BA in translation/interpreting vs other type of BA), source text domain (financial vs legal) and type of task (in-class activity vs examination). In a second step, we examine the error categories qualitatively with a view to determining whether they are similar or different in translation and post-editing.

We fitted a mixed-effects regression model with number of errors as response variable, student_id as random variable, and translation_method, translation_experience, domain, and task_type as fixed variables. The model indicates, among other things, that translation method has a significant impact on quality, with student post-edited texts containing nearly twice as many errors as translations from scratch. This trend is mainly due to the fact that students fail to identify about a third of the MT errors (c. 60% of the erroneous segments in the final PE products are MT errors left undetected). This finding supports the idea that students need to be trained in MT error detection, a skill specific to PE. Our data further show that when students identify errors in the MT, they are able to correct them successfully in the vast majority of cases (c. 90%). In translation, the most frequent error categories are Content, Lexis & Terminology, and Register & Style. In post-editing, we find the same categories, in roughly the same proportions, in errors introduced by students while post-editing (NE-INTR). This suggests that students make similar mistakes when translating and post-editing. Among MT errors that went undetected and are thus still present in the final PE products (E-MISS), Lexis & Terminology rank first, followed by Content, and Register & Style, which once again shows that these are the most problematic areas for our students.

In our presentation, we will outline the main findings of our study and discuss their implications for translator education.

References

Bowker, L., & Bennison, P. (2003). Student Translation Archive: Design, development and application. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in Translator Education* (pp. 103–117). Routledge.

Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N., & Volanschi, A. (2011). Designing a learner translator corpus for training purposes. In N. Kübler (Ed.), *Corpora, Language, Teaching, and Resources: From Theory to Practice* (pp. 221–248). Peter Lang.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2017). Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta*, *62*(2), 245–270.

Garcia, I. (2010). Is machine translation ready yet? *Target*, *22*(1), 7–21.

Granger, S., & Lefer, M.-A. (2023). Learner translation corpora: Bridging the gap between learner corpus research and corpus-based translation studies. *International Journal of Learner Corpus Research*, *9*(1), 1–28.

Granger, S., & Lefer, M.-A. (2021). *Translation-oriented Annotation System manual (Version 2.0)*. *CECL Papers 3*. Centre for English Corpus Linguistics/UCLouvain. https://uclouvain.be/en/research-institutes/ilc/cecl/cecl-papers.html

Granger, S., & Lefer, M.-A. (2020). The Multilingual Student Translation corpus: a resource for translation teaching and research. *Language Resources and Evaluation*, *54*, 1183–1199.

Kübler, N., Mestivier, A., & Pecman, M. (2022). Using comparable corpora for translating and post-editing complex noun phrases in specialized texts: Insights from English-to-French specialized translation. In S. Granger, & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-based Translation Studies* (pp. 237–266). Bloomsbury.

Kutuzov, A., & Kunilovskaya, M. (2014). Russian Learner Translator Corpus: Design, Research Potential and Applications. In P. Sojka, A. Horák, I. Kopeček, & K. Palak (Eds.), *Text, Speech and Dialogue* (pp. 315–323). Lecture Notes in Computer Science. Springer.

Lefer, M.-A., Bodart, R., Obrusník, A., & Piette, J. (2023). The Post-Edit Me! project. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.

Lefer, M.-A., Piette, J., & Bodart, R. (2022). *MTPEAS manual: Machine Translation Post-Editing Annotation System*. OER-UCLouvain. https://oer.uclouvain.be/jspui/handle/20.500.12279/829

Nitzke, J., Hansen-Schirra, S., & Canfora, C. (2019). Risk Management and Post-editing Competence. *JosTrans*, *31*, 239–259.

Uzar, R. S. (2002). A corpus methodology for analysing translation. In S.E.O. Tagnin (Ed.), *Cadernos de Tradução: Corpora e Tradução* (pp. 235–263). NUT, 1(9).

A Contrastive Corpus Approach to Absence: Methodological Developments and Theoretical Implications

Antonina Bondarenko

Université Paris Cité (Clillac-Arp) & INALCO (SeDyL), France

KEYWORDS: verbless sentences, contrastive corpus methodology, semantico-pragmatics, sentential models

This paper focuses on the controversial and often marginalised phenomenon of the verbless sentence, i.e. structures in which the typical syntactic marker of sentential status – the verbal predicate – is absent. Although the structures exist in many languages (e.g. Benveniste, 1971), the difficulty of automatically processing them in corpora has greatly limited their study. Presenting recent developments in the corpus treatment of the structures, the paper explores how the contrastive corpus approach has enhanced our understanding of the verbless phenomenon and its implications for language models.

We examine the phenomenon in English and Russian, two languages that have profoundly different typological characteristics with regard to the verbless issue. Notably, Russian is known for the most liberal use of verbless structures among the Indo-European languages, a developed morphological case system that contributes to a remarkable capacity for syntactic ellipsis, the unmarked absence of a copula in the present tense, e.g. (1), as well as the capacity to easily omit full lexical verbs without requiring an antecedent, e.g. (2), while English is typically associated with a dependence on the finite verb phrase (Stassen, 2013; McShane, 2000).

(1)	*Я*	Мария.		(2)	*Я*	в	магазин.
	ja	*Marija*			*ja*	*v*	*magazyn*
	I	*Marija*			*I*	*to*	*store*

In order to (a) provide a corpus-based description of the semantico-pragmatic features associated with the absence of the verb in English and Russian and (b) explore the implications of the results for existing theoretical perspectives of the 'sentence', we develop a multi-dimensional methodological framework that combines the contrastive approach of Guillemin-Flescher (2003), i.e. analysing recurring translation patterns for language-specific typological regularities, with corpus-driven methods, statistical tools and translation studies considerations.

We started by lifting the barrier on the automatic processing of open grammatical structures centered on absence; developing a method for the automatic retrieval of verbless structures, as well as their multiple translation correspondences and their contexts (Bondarenko, 2021; 2019). Potential pitfalls of using parallel-text studies for linguistic conclusions (raised by e.g. Nádvorníková, 2017; McEnery & Xiao, 2008; Stolz, 2007; Malmkjaer, 1998) were addressed in the corpus design and multidirectional analysis. Notably, we built a 1,4-million-word parallel and comparable corpus of 19th–21st century English and Russian discourse-based fiction, morpho-syntactically tagged and fully sentence-aligned across multiple translations (Bondarenko, 2021) which was then analyzed from three perspectives. From a *monolingual* perspective, translations were treated as genuine language samples (following Zanettin, 2013; Olohan, 2002; Baker, 1993; Biber, 1993) and compared with originals in terms of a specificity analysis of the constituents that make up verbless sentences (including the calculation of statistically-key forms, lemmas, morphosyntactic categories and n-grams, against a reference corpus, and their semantico-pragmatic classification). Secondly, we looked for *reciprocal* patterns across (a) multiple translations, (b) texts and (c) translation-directions, paying attention particularly to *verbal* correspondences and their correlation with manually annotated syntactic ellipsis, information structure and speech acts criteria. A *third-language* sub-corpus of Russian and English translations from French was included to control for source language interference.

Through contrastive corpus analysis we contribute new evidence that verbless structures are *not syntactic reductions*, i.e. supporting arguments in both recent minimalism and construction grammar (e.g. Barton &

Progovac, 2005; Elugardo & Stainton, 2005; Goldberg & Perek, 2019) against the existence of supplementary hidden verbal syntactic structure. Our results show that syntactic antecedent-based ellipses are significantly overrepresented not in Russian, the more elliptically-productive language, but rather in English (X^2 p<2.20E-16). Syntactic explanations are insufficient to account for the frequency variation between Russian and English.

Contrastive evidence also leads us to argue that verbless structures are *not semantic reductions*. Verbs that are pragmatically implicated by verbless sentences were found to be a part of the information-structural focus (in the sense of Lambrecht, 1994), which by definition is unpredictable and unreconstructible, e.g. (3).

(3) [Ну, а здесь ничего, здесь нет монастырских жен, а *монахов* штук двести. Честно. **Постники.** Сознаюсь.] [Well, there's nothing like that here, no monastery wives, and about two hundred *monks.* It's honest. **They fast**. I admit it.]

Постники.	They fast.
postniki	
NN.MPL.NOM	
people_who_fast	

This finding challenges the extent to which the verbless phenomenon can be accounted for in terms of the omission of a predictable and semantically reconstructable element. Furthermore, it provides evidence that the bare constituents of a verbless sentence can be sufficient to constitute full instances of predication.

The contrastive corpus method also revealed *pragmatic explanations* as to why Russian has twice as many verbless sentences as English. Rather than syntactic or semantic recoverability, we identify differences concerning (a) topic activation (English tends to re-active the topic which correlates with the realization of the verb), and (b) the marking of the difference between direct and indirect speech acts in questions, e.g. (4).

(4) [Excommunication is an argument in a debate. The speaker interrupts his interlocutor and utters two questions, that immediately follow one another.]

Как	это	отлучение,	что	за	отлучение?
kak	*èto*	*otlučenie*	*čto*	*za*	*otlučenie*
what	*this*	*excommunication*	*what*	*particle*	*excommunication*

What do you mean – excommunication? What excommunication?

Furthermore, semantico-pragmatic classification of statistically key elements reveals additional differences between the languages in the use of the structures. Indefinite reference and wh-words were found to be key in English verbless sentences, e.g. (5), while for Russian, emphasis of intensity is particularly important, e.g. (6). Finally, we also draw attention to the importance of genre differences for the frequency of this phenomenon.

(5) [He moved his eyes quickly away. **What a strange meeting on a strange night.** He remembered nothing like it save one afternoon a year ago when he had met an old man in the park and they had talked…]

What a strange meeting on a strange night.

(6) [Одного только не забывайте – брандмейстер принадлежит к числу самых опасных врагов истины и свободы, к тупому и равнодушному стаду нашего большинства. **О, эта ужасная тирания большинства!** Мы с Битти поем разные песни.]

О,	эта	ужасная	тирания	большинства!
o,	*èta*	*užasnaja*	*tiranija bol'šenstva*	
INTJ	DEM.F	ADJ.NOM	NN.NOM	NN.GEN
o	*this*	*terrible tyranny*	*of_majority*	

The results thus revealed through contrastive corpus methodology give us new empirical grounds for defending the sentential status of the theoretically controversial phenomenon of the verbless utterance. They lead us toward a model of the sentence at the syntactic-semantic-pragmatic interface that would account for the syntactically non-canonical structures and emphasize the communicative functions of language.

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP & Translation: Studies in Language Engineering in Honour of Juan C. Sager* (pp. 175–186). John Benjamins Publishing Company. https://doi.org/10.1075/btl.18.17bak

Barton, E., & Progovac, L. (2005). Nonsententials in Minimalism. In R. Elugardo & R. J. Stainton (Eds.), *Ellipsis and Non-Sentential Speech* (pp. 71–94). Springer. https://doi.org/10.1007/1-4020-2301-4_4

Benveniste, É. (1971). *Problems in General Linguistics* (M. E. Meek, Trans.). University of Miami Press. (Original work 'Problèmes de Linguistique Générale' published 1966)

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Bondarenko, A. (2019). A corpus-based contrastive study of verbless sentences: Quantitative and qualitative perspectives. *Studia Neophilologica*, *91*(2), 175–198. https://doi.org/10.1080/00393274.2019.1616221

Bondarenko, A. (2021). *Verbless and Zero-Predicate Sentences: An English and Russian Contrastive Corpus Study* [Doctoral Dissertation, Université de Paris]. https://theses.hal.science/tel-03706573

Elugardo, R., & Stainton, R. J. (Eds.). (2005). *Ellipsis and Non-Sentential Speech*. Springer. https://doi.org/10.1007/1-4020-2301-4

Goldberg, A., & Perek, F. (2019). Ellipsis in construction grammar. In J. Van Craenenbroeck & T. Temmerman (Eds.), *The Oxford Handbook of Ellipsis* (pp. 1–21). Oxford Handbooks Online. https://doi.org/10.1093/oxfordhb/9780198712398.013.8

Guillemin-Flescher, J. (2003). Théoriser la traduction. *Revue Française de Linguistique Appliquée*, *8*(2), 7–18. https://doi.org/10.3917/rfla.082.07

Lambrecht, K. (1994). Information Structure and Sentence Form: Topic, Focus and the Mental Representation of Discourse Referents. Cambridge University Press. https://doi.org/10.1017/CBO9780511620607

Malmkjaer, K. (1998). Love thy neighbour: Will parallel corpora endear linguists to translators? *Meta*, *43*(4), 534–541. https://doi.org/10.7202/003545ar

McEnery, A., & Xiao, R. (2008). Parallel and comparable corpora: What is happening? In G. Anderman & M. Rogers (Eds.), *Incorporating Corpora: The Linguist and the Translator* (pp. 18–31). Multilingual Matters. https://doi.org/10.21832/9781853599873-005

McShane, M. (2000). Verbal ellipsis in Russian, Polish and Czech. *The Slavic and East European Journal*, *44*(2), 195–233. https://doi.org/10.2307/309950

Nádvorníková, O. (2017). Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela [Online] HS*, *21*, 1–28. https://doi.org/10.4000/corela.4810

Olohan, M. (2002). Corpus linguistics and translation studies: Interaction and reaction. *Linguistica Antverpiensia*, *1*, 419–429. https://doi.org/10.52034/lanstts.v1i.29

Stassen, L. (2013). Zero copula for predicate nominals. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/120

Stolz, T. (2007). Harry Potter meets Le Petit Prince: On the usefulness of parallel corpora in crosslinguistic investigations. *STUF - Sprachtypologie Und Universalienforschung: Language Typology and Universals*, *60*(2), 100–117. https://doi.org/10.1524/stuf.2007.60.2.100

Zanettin, F. (2014). Corpora in translation. In J. House (Ed.), *Translation: A Multidisciplinary Approach* (pp. 178–199). Palgrave Macmillan. https://doi.org/10.1057/9781137025487_10

Who writes more plainly? A case study comparing plain language summaries produced by science writers and journal article authors

Lynne Bowker

University of Ottawa, Canada & NAWA Visiting Researcher at Adam Mickiewicz University, Poznań, Poland

KEYWORDS: interlingual translation, plain language, science communication, scholarly publishing, Canadian Science Publishing

In the scholarly community, there is a growing movement to make specialized knowledge available to non-expert communities (e.g. UNESCO, 2021; Science Europe, 2022; Stoll et al., 2022). In response, some academic journals now encourage the production of plain language summaries (e.g. SAGE 2023; Taylor & Francis 2023). A plain language summary is a type of constrained writing that could be described as an intralingual translation between specialized and general language. Translation scholars such as Whyatt (2017) and Kotze (2022) argue that we need to pay more attention to both intralingual translation and constrained language varieties beyond traditional translation.

The Canadian Science Publishing (CSP) group has a collection of 23 journals across a range of scientific fields. Since 2015, CSP has adopted a two-pronged approach to producing plain language summaries of their journal content. Firstly, like other academic publishers, CSP encourages article authors (i.e. subject experts) to produce their own plain language summaries, and these are hosted on CSP's page on the open online publishing platform Medium (CSP, 2023). Secondly, unlike other publishers, CSP also commissions science writers to produce plain language summaries, which are hosted on the CSP's blog. As of the end of December 2022, there were 96 plain language summaries by science writers and 216 plain language summaries by article authors.

Previous researchers (e.g. Stricker et al., 2020) have compared plain language summaries to conventional scientific abstracts and have established that the summaries score better for various readability measures. However, to our knowledge, no previous research compares plain language summaries written by science writers to those written by article authors. CSP therefore presents an interesting opportunity for a case study comparing the features of plain language summaries prepared by these two groups.

We conducted a pilot study using the CSP collection. For the pilot study, we created a corpus containing 50 plain language summaries by science writers that were based on research from one of CSP's journals for a corpus totalling 32,322 words. Next, we created a corpus of author-produced plain language summaries containing 100 summaries totalling 32,733 words. (Note that since the author-produced summaries were shorter than the science writer summaries, we adjusted the number to achieve a comparable corpus size). We compared the two corpora using simple measures corresponding to CSP's guidelines for writing plainly (e.g. avoid passives, use short sentences). The measures include average text length, average number of sentences per paragraph, average number of words per sentence, percentage of passives, and readability scores (Flesch Reading Ease and Flesch-Kincaid Grade Level). Overall, the texts produced by science writers more closely adhere to the guidelines for writing in plain language; however, the difference is less remarkable than one might have expected, given that the science writers have explicit training in writing or journalism (as indicated in their biographies). This raises questions such as whether a minimal amount of training and improved guidelines could close the gap between the two groups, whether there is value in having both groups write plain language summaries, or whether either group should be pushed to go further in their efforts to write plainly.

References

Kotze, H. (2022). Translation as constrained communication: principles, concepts and methods. In S. Granger & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-based Translation Studies* (pp. 67–97). Bloomsbury.

Science Europe. 2022. Position statement: Science communication for greater research impact. DOI: 10.5281/zenodo.6645074 https://zenodo.org/record/6645075

Stoll, M., Kerwer, M., Lieb, K. & Chasiotis, A. (2022). Plain language summaries: A systematic review of theory, guidelines and empirical research. *PLOS One*, 17(6), Article e0268789. https://doi.org/10.1371/journal.pone.0268789

Stricker, J., Chasiotis, A., Kerwer, M. & Günther, A. (2020). Scientific abstracts and plain language summaries in psychology: A comparison based on readability indices. *PLoS ONE, 15(4),* Article e0231160. https://doi.org/10.1371/journal.pone.0231160

UNESCO. 2021. Recommendation on Open Science. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

Whyatt, B. (2017). Intralingual translation. In A. Ferreira & J. Schwieter (Eds.), *The Handbook of Translation and Cognition* (pp. 176–192). Wiley-Blackwell.

Publisher sites

Canadian Science Publishing (CSP). 2023. https://cdnsciencepub.com/authors-and-reviewers/writing-a-plain-language-summary

SAGE. 2023. https://us.sagepub.com/en-us/nam/plain-language-summaries

Taylor & Francis 2023. https://authorservices.taylorandfrancis.com/publishing-your-research/writing-your-paper/how-to-write-a-plain-language-summary/#

PINC for everyone. Introducing Polish Interpreting Corpus

Agnieszka Chmiel[1], Danijel Korzinek[2], Marta Kajzer-Wietrzny[1], Przemysław Janikowski[3]

[1]Adam Mickiewicz University, Poznan, Poland, [2]Polsko-Japońska Akademia Technik Komputerowych, [3]University of Silesia, Poland

KEYWORDS: interpreting corpus, speech corpus, parallel corpus, web interface, European Parliament

In this presentation, we introduce PINC — the Polish Interpreting Corpus compiled as a part of the project "Extreme language control: activation and inhibition as bilingual control mechanisms in conference interpreting" funded by the National Science Centre (2019-2024) (MO-2018/30/E/HS2/00035 (SONATA BIS 8). It is a Polish-English and English-Polish corpus of short European Parliament speeches and their interpretations. The Polish Interpreting Corpus (PINC) adds to an ever-growing family of interpreting or intermodal corpora derived from the European Parliament debates called by Bernardini et al. (2018) "the EPIC suite of corpora". As summarised by Bernardini et al. (2018: 22), "[t]he availability of interpretations and translations from and into a large number of languages, the ease of access to the videos (downloadable from the Internet), and the high professional standards of the interpreters" makes this source very promising for interpreting corpora, which is why the EPIC suite is constantly growing".

Several aspects of the PINC compilation process have been modelled on the work of the creators of the other corpora of the EPIC suite (e.g. EPIC or EPTIC). Thus, the texts compiled in the PINC corpus follow the same topic classification as EPIC and EPTIC for ease of comparison; similar contextual metadata has been collected and transcription guidelines were, to a large extent, very much alike.

The uniqueness of PINC consists in the specific language combination, careful balancing of the mode of delivery, detailed interpreter voice identification (Korzinek 2020a, Korzinek 2021), speech-to-text (Korzinek 2019, 2020b) and sentence alignment of the whole corpus and in the specific tools employed to automatise parts of the compilation process. Word-level ST and TT alignment was automatically performed in SimAlign and manually corrected for all ST–TT equivalents. Since we are interested in a fully bidirectional analysis, PINC consists of four balanced subcorpora: Polish source texts (ST-PL), their interpretations into English (TT-EN), English source texts (ST-EN) and their interpretations into Polish (TT-PL). All of these were collected from the Europarl website from plenary session recordings of the European Parliament sittings taking place between January 2009 and September 2010.

The PINC corpus comprises texts ranging between 100 and 500 words. Including a higher number of shorter speeches rather than fewer longer ones made it possible to achieve greater variation within the data. PINC is annotated for mode of delivery distinguishing between: impromptu (unscripted) and read-out (scripted) speeches. In the course of compilation of PINC, a decision was made not to include speeches of varying degree of scriptedness.

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

Interpreters are key in any interpreting corpus. Professionals working during the European Parliament plenary sessions are carefully selected in a process designed to guarantee top quality interpreting services at the EU institutions. The usual problem with EP data, however, is that the only detail allowing us to distinguish between them is their voice. Controlling for individual variation is desired in many empirical

studies, hence PINC does include precise metadata on interpreter identity. This will greatly enhance the control of the individual variation in further analyses. In our presentation, we will share insights from the process of interpreter identification using a two-step procedure involving manual/ human identification and automatic identification.

In the course of our presentation, we will also show the corpus web interface and we will briefly describe a few studies showcasing the use of such a corpus focusing on fluency parameters (Chmiel et all 2022), such as speaking rate and pauses in interpretations, spillover effect in interpreting (Chmiel et al 2023), factors modulating EVS, grammatical features of interpretese or the use of collocations by interpreters working into A and B language.

References

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, B. & Defrancq, B. (eds.), *Making way in corpus-based interpreting studies*, 21–42. Singapore: Springer. DOI: 10.1007/978-981-10-6199-8_2.

Chmiel, A., P., Koržinek, Janikowski, D., Lijewska, A., Kajzer-Wietrzny, M., Jakubowski, D. & Polakowska, D (2022). Fluency parameters in the Polish Interpreting Corpus (PINC). in Kajzer-Wietrzny, M., Ferraresi, A., Ivaska, I.& Bernardini, S. (eds.), *Mediated discourse at the European Parliament: Empirical investigations*, 63–91. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977042

Chmiel, A., Janikowski, P., Koržinek, D., Lijewska, A., Kajzer-Wietrzny, M., Jakubowski, D. & Plevoets, K. (2023) Lexical frequency modulates current cognitive load, but triggers no spillover effect in interpreting, *Perspectives*, DOI: 10.1080/0907676X.2023.2218553

Koržinek, D. (2019). Corrector. https://github.com/danijel3/Corrector.

Koržinek, D. (2020a). Speaker identification. PINC Project.

Koržinek, D. (2020b). Automating word segmentation. PINC Project. https://pincproject2020.wordpress.com/2020/04/08/automating-word-segmentation/https:/pincproject2020.wordpress.com/2020/04/28/speaker-identification/

Koržinek, D. & Chmiel, A. 2021. Interpreter identification in the Polish Interpreting Corpus. *Revista Tradumàtica* 19. 276–288.

Different translators, different translation solutions? Analyzing n-grams in a French-to-Dutch multiple-translation corpus – and what it reveals about expertise, priming and cognitive routinisation

Gert De Sutter[1], Marie-Aude Lefer[2] & Koen Plevoets[1]

[1]Ghent University, Belgium, [2]UCLouvain, Belgium

KEYWORDS: translation variance, expertise, priming, cognitive routinization, controlled corpus design

Emotions are multi-faceted phenomena, including physiological, behavioral, cognitive and experiential aspects. They are rooted in evolutionary old brain systems and serve to mobilize the organism for fast responses in survival-relevant situations. Emotions can be described by words which themselves can arouse emotions, apparently linking evolutionary more recent linguistic cognitive systems with older, more rudimentary ones, thereby altering word processing. Processing of emotionally arousing words has been shown to differ from neutral ones: Emotional words are responded to faster, induce stronger pre-response activation of the motor systems (Kissler & Koessler, 2011), elicit higher amplitude event-related potentials at several processing stages (Kissler & Herbert, 2013) and hemodynamic responses in deep "emotional brain" nuclei such as the amygdala (Herbert et al., 2009), and they are better remembered than neutral words. This corroborates the view that processing emotional words is embodied. Yet, recent evidence has shown that what sets emotional words apart from neutral ones is not fixed but malleable by social context (Schindler et al., 2019a; Schindler et al., 2019b]. Moreover, recent evidence showed that several of the physiological markers thought to index activation of emotional brain regions during emotional word processing persist even in the absence of those regions and that subjective appraisal and emotional memory need not be altered by loss of medial temporal lobe structures (Kissler et al., submitted). Thus, these data paint a complex picture on the embodiment of emotional words: on the one hand, a large body of data shows that the processing of these words is often draws on body-related brain structures. On the other hand, this embodiment is itself contextualized and the processing system seems quite robust against considerable damage to emotion-related brain structures. Therefore, data suggest that representation and processing of emotional words are realized in multiple, interactive and partly redundant systems that while typically embodied and situated, can show a considerable degree of autonomy.

References

Castagnoli, S. (2020). Translation choices compared: Investigating variation in a learner translation corpus. In S. Granger, & M.-A. Lefer (Eds.), *Translating and Comparing Languages: Corpus-based Insights*. *Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference. Corpora and Language in Use Proceedings 6* (pp. 25–44). Presses Universitaires de Louvain.

Castagnoli, S. (2023). Exploring variation in student translation. *International Journal of Learner Corpus Research*, *9*(1), 97–125.

Diessel, H. (2019). *The grammar network*. Cambridge University Press.

Halverson, S. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M. Lefer, & I. Delaere (Eds.), *Empirical Translation Studies: New Methodological and Theoretical Traditions* (pp. 9–46). De Gruyter.

Hartsuiker, R. J., Beerts, S., Loncke, M., Desmet, T., & Bernolet, S. (2016). Cross-linguistic structural priming in multilinguals: Further evidence for shared syntax. *Journal of Memory and Language, 90*, 14–30.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M. Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V (2014). The Sketch Engine: ten years on. *Lexicography*, *1*, 7–36.

Lefer, M.-A. (2020). Parallel corpora. In M. Paquot, & S. Th. Gries, *Practical Handbook of Corpus Linguistics* (pp. 257–282). Springer.

Malmkjær, K. (1998). Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators? *Meta: Translators' Journal, 43*(4), 534–541.

Prinzie, T. (2022). Variability in a multiple translation corpus as evidence for cognitive processes. Paper delivered at the *CogLing Days (Belgium Netherlands Cognitive Linguistics Association): Language, Discourse and Cognition*.

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

A discourse marker in interpreter-mediated police interviewing with drafting. A corpus-based approach to dialogue interpreting

Bart Defrancq

Ghent University, Belgium

KEYWORDS: dialogue interpreting, police interpreting, discourse marker, interaction

Corpus-based interpreting studies has rapidly evolved over the last ten to fifteen years. The availability of corpus materials from several international institutions and the significant improvement of automatic transcription tools removed two of the main obstacles to research that Shlesinger (1998) identified. Important, though comparatively small-scale, corpora have been collected in various research centres. Quite a few research papers applying corpus-based methods have been published over the years, focusing on lexical and pragmatic properties of interpreted texts, translation universals and cognitive load, and on a small number of language pairs. However, these studies have focused on conference interpreting. Dialogue interpreting has mostly been overlooked. It is remarkable, for instance, that the special issue of the *Interpreters' Newsletter* (issue 22), published in 2017 and specifically devoted to *Corpus-based Dialogue Interpreting Studies*, as its title reads, contains not a single contribution showcasing empirical work based on a real corpus (Bendazzoli 2017).

As a result of this, usage patterns widely observed in conference interpreting cannot be generalised for interpreting per se, as we lack sufficient data of other types of interpreting to make that case. The untriggered use of particular discourse markers in simultaneous conference interpreting, for instance, is widely attested in several language combinations (Defrancq et al. (2015) for French-English/Dutch; Götz 2020 for English-Hungarian, and Gumul & Bartłomiejczyk 2022 for English-Polish) and has been associated with the high cognitive load interpreters experience (Gumul 2021). To find out whether this is really a trait of interpreting in general or rather of simultaneous interpreting, large-scale corpus studies of other kinds of interpreting are required.

Our study reports on a systematic analysis of the use of one particular Dutch discourse marker in a corpus of police interpreting (IMPID). The corpus includes interpretations by nine different interpreters recorded in twelve police interviews and comprises approximately 114,000 tokens of interpretation on a total token count of 245,000 for the entire corpus. Interpretation is carried out bidirectionally between Dutch and seven other languages: Arabic, English, French, Greek, Pashto, Romanian, and Turkish. The nature of the cases varied: half of the interviews were recorded as part of sham marriage procedures, while the other half included typical criminal cases, such as drug dealing, assault, human trafficking and possession of a prohibited weapon. The Dutch discourse marker *dus* ('so') was selected based on its observed high frequency in simultaneous interpretation and, in particular, on the large share of untriggered occurrences (Defrancq et al. 2015). The aim of the study was to verify whether other types of interpreting show the same tendencies as simultaneous interpreting. The use of equivalents of *dus* in the other languages included in the corpus was beyond the scope of the study.

The main results can be summarised as follows: in terms of frequencies, interpreters do not divert from general usage patterns for spoken Dutch, but individual variation across interpreters is high, in so far that only two interpreters produce roughly half of the occurrences. Variation could not be associated with interpreter- or case-related variables. Most importantly, interpreters' use of *dus* is mostly disconnected from the speech they are interpreting. Nearly 90% of the 265 occurrences have no equivalent in the corresponding source utterances. This general pattern seems to hold for all interpreters. While explicitation seems to be at play in a certain number of cases, the bulk of instances serves interaction coordination purposes (turn taking and turn yielding). This was expected in a context of dialogue interpreting where interpreters coordinate speech (Wadensjö 1998). Interestingly, interpreters use a Dutch particle to take the floor from the interviewee who does not speak Dutch. In many cases, the turn-taking fails and interpreters, as a result, are seen to engage in a stretch of simultaneous interpreting. Given the

cognitive challenges interpreting poses it is not surprising to also find a substantial number of filler instances where interpreters struggle to understand the source speech or to articulate their interpretation. *Dus* is usually part of a string of several different hesitation markers in those cases. Some interesting cases of so-called discursive control (Ainsworth 2018) enforced by *dus* were also observed, i.e. instances where *dus* is used to point the police officer to the most recordable segment of the information provided by the interviewee. This further confirms the special relationship interpreting holds with drafting of written records during the interview (Pöchhacker & Kolb 2009).

References

Ainsworth, J. (2018). Anatomy of a false confession: the linguistic and pyschological characteristics of a false confession. In G. Tessuto, V. Bhatia, & J. Engberg (Eds), *Frameworks for Discursive Actions and Practices of the Law* (pp. 23-39). Newcastle upon Tyne: Cambridge Scholars.

Bendazzoli, C. *(*2017). Editorial. A dialogue on dialogue interpreting (DI) corpora. *The Interpreters' Newsletter, 22*, VII-XVII.

Defrancq, B., Plevoets, K., & Magnifico, C. (2015). Connective markers in interpreting and translation: where do they come from? In J. Romero Trillo (Ed), *Corpus pragmatics in translation and contrastive studies* (pp. 195-222). Singapore: Springer.

Götz, A. (2020). Discourse markers and connectives in interpreted Hungarian discourse: A corpus-based investigation of discourse properties and their interdependence. *Beszédtudomány – Speech Science* 2020, 259–284.

Gumul, E. (2021). Explicitation and cognitive load in simultaneous interpreting Product- and process-oriented analysis of trainee interpreters' outputs. *Interpreting*, *23*(1), 45-75.

Gumul, E., & Bartłomiejczyk, M. (2022). Interpreters' explicitating styles: A corpus study of material from the European Parliament. *Interpreting*, *24*(2), 164-191.

Pöchhacker, F., & Kolb, W. (2009). Interpreting for the record: A case study of asylum review hearings. In S. Hale, U. Ozolins, & L. Stern (Eds), *Quality in Interpreting – A Shared Responsibility* (pp. 119-134). Amsterdam: Benjamins.

Shlesinger, M. (1998). Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies. *Meta*, *43*(4), 486-493.

Wadensjö, C. (1998). *Interpreting as Interaction*. London: Longman.

Corpus, corpus on the wall, who is the most collocational of them all? A comparison of dependency-defined collocations in learner translations and learner writing

Adriano Ferraresi, Maja Miličević Petrović, Silvia Bernardini

University of Bologna, Italy

KEYWORDS: collocations, constrained language, Gravitational Pull Hypothesis, learner writing, learner translation

In this paper we investigate the use of collocations by Italian-speaking translation students at Master's level translating into and writing in English as a second language (L2). Both translation and learner production are known to diverge from non-mediated texts produced by native speakers of a language on several dimensions, including the phraseological one, whereby both translation and L2 writing are characterized by highly frequent rather than strongly associated combinations (Kenny, 2001 and Durrant & Schmitt, 2009, among others). A growing body of work is exploring the added value of *learner* translation data in studies of mediated language production (Granger & Lefer, 2023), but to the best of our knowledge no attempt has been made to compare learners' writing and translation directly.

We construe the two tasks – L2 writing and translation by learners – as different modes of constrained language production (Kruger & van Rooy, 2016; Kotze 2020), and compare them adopting as an additional theoretical framework the Revised Gravitational Pull Hypothesis (RGPH; Halverson, 2017). The two tasks share the constraint given by bilingual activation, with the first language influencing the L2 in both cases, and by a semi-expert level of proficiency. They instead differ on the text production constraint, which is independent/unmediated in L2 writing and dependent/mediated in translations (cf. Kotze, 2020, p. 346-347). In considering the RGPH, we primarily focus on second/target language salience ("magnetism"), considering also the relevance of first/source language prominence ("gravitational pull") and interlingual links ("connectivity") – proposed as key factors in the translation process – for L2 writing. We specifically aim to establish: (1) whether texts written directly in English and translations into English differ in their level of "collocationality", roughly defined as the presence of strongly associated English lexical combinations (magnetism); (2) whether Italian equivalents of English collocations used in translations and essays are more/less/equally strong across the two modes (gravitational pull); (3) whether collocations across the two modes are more/less/equally likely to result from the existence of direct cross-linguistic links between English and Italian (connectivity).

We attempt to answer these research questions using two corpora that were assembled opportunistically gathering learner production from a population of students enrolled in the same translation MA. In the attempt to favor representativeness and ecological validity of the sample over balance and comparability, the two datasets vary in terms of size and structure. Our writing corpus comprises 106 student essays on 13 topics from the domain of corpus linguistics, with each essay written by a different student (total words: 224,968), while the translation corpus contains 131 translations based on 37 Italian source texts, with 19 contributing students and topics coming from multiple specialised domains such as psychology, marketing and economics (total words: 27,393). The collocations we study are word pairs extracted using syntactic dependencies as codified by the spaCy + UDPipe Universal Dependency Parser (Straka, 2018; TakeLab, 2021). Sixteen dependency relations are captured that involve adjectives, nouns, verbs, or adverbs and enable the extraction of collocational patterns typical of English (e.g. *nmod* => noun pre-modification by another noun, as in *time period*). Collocations are extracted from the two corpora and their collocational strength is assessed by calculating two association measures (AM), namely mutual information and logDice, following an initial filtering-out of very rare and highly specialized collocations. Italian equivalents of a sub-set of the extracted collocations are searched for drawing on lexicographic evidence and machine translation, and assigned AM scores following the same method as their English counterparts.

The initial analysis we performed adopts a univariate design, based on bootstrapped independent sample *t*-tests with 10,000 replications using the BCa method (cf. LaFlair et al., 2015, p. 51) and aggregating data at the level of individual learners. The analysis shows that lexical association scores are significantly higher in translations than in L2 writing (for MI-based collocations: t=−8.85; 95% BCa C.I.: −11.826, −6.356; for log-dice-based collocations: t=−3.45; 95% BCa C.I.: −6.01, −0.98), leading us to conclude that second/target language salience effects are more visible in translation than in independent L2 writing. Preliminary qualitative analyses of a subset of collocations shared and unique to either mode shows that translations also feature more collocations with direct cross-linguistic links, as evidenced by the number of word pairs attested in bilingual dictionaries, while source/first language seems to affect both modes similarly, as shown by the AM scores of the manually identified Italian equivalents.

In the conference paper, we will report on additional quantitative explorations attempting to address two of the problems faced in the initial analysis: (1) Our conceptualisation of second/target language salience was through collocationality levels of the L2/translated texts. In light of the RGPH, would it be feasible to identify a measure of how different the corpora are from a native unmediated baseline, with L2 collocationality level, the collocational status of first/source language equivalents and connectivity as factors included in a multifactorial analysis (e.g. along the lines of Lefer & de Sutter, 2022)? (2) How can first language equivalents and connectivity be reliably established in learner writing – where a source text is not present –, and in this light what are the potential affordances and limitations of the RGPH as a more general model of production by bilinguals? Finding answers to these questions would allow us to take steps toward setting up a methodological framework for the analysis of phraseological aspects of (learners') constrained communication.

References

Lefer, M.-A. & De Sutter, G. (2022). Using the Gravitational Pull Hypothesis to explain patterns in interpreting and translation: The case of concatenated nouns in mediated European Parliament discourse. In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska & S. Bernardini (Eds.), *Mediated discourse at the European Parliament: Empirical investigations* (pp. 133–159). Language Science Press.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics*, *47*(2), 157–177.

Granger, S. & Lefer, M.-A. 2023. Learner translation corpora: Bridging the gap between learner corpus research and corpus-based translation studies. *International Journal of Learner Corpus Research*, *9*(1), 1–28.

Halverson, S. L. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds), *Empirical translation studies: New theoretical and methodological traditions* (pp. 9–46). De Gruyter.

Kenny, D. (2001). Lexis and creativity in translation. A corpus-based approach. St. Jerome.

Kotze, H. (2020). Converging what and how to find out why. In L. Vandevoorde, J. Dams, & B. Defrancq (Eds.), *New empirical perspectives on translation and interpreting* (pp. 333-71). Routledge.

Kruger, H., & van Rooy, B. (2016). Constrained language. A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, *37*(1), 26–57.

LaFlair, G. T., Egbert, J, & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46–77). Routledge.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

TakeLab (2021). SpaCy + UDPipe. Available at https://github.com/TakeLab/spacy-udpipe.

Ecological validity in corpus-based and experimental translation research

Jonas Freiwald, Zoe Miljanovic, Stella Neumann

RWTH Aachen University, Germany

KEYWORDS: translation process research, corpus linguistics, ecological validity, grammar, inferential statistics

This study provides insight into the relationship between corpus data and experimental data by comparing the translations produced under three different conditions: in a translation research lab, in the translators' usual working environment, and included as part of a corpus. The results will help us to explain some of the discrepancies between findings based on corpus data and lab experiments and to test the ecological validity of experimental results in translation process research (TPR).

TPR has become more popular over the last decades and experimental rigour in translation experiments has been steadily increasing. However, this data is collected outside the usual working environment of translators and the observed phenomena may be caused by the experimental context and may not represent authentic translation behaviour. The source texts in translation experiments can have unusual characteristics due to experimental manipulation (Schaeffer and Carl 2017, 118) or other constraints imposed by the experimental setup. Participants are not always allowed to use external resources (Schaeffer and Carl 2017) and so research on a topic or even simple dictionary look-ups that are typically part of a translator's working routine are not possible. Lab experiments often use context-free sentences, which have been shown to be translated differently from longer texts while paragraph-sized texts were considerably more similar to the full-text condition (Heilmann et al. 2019). The use of language technologies like translation memories and workbenches is common practice in translations, but the typical keystroke logging and eye tracking experiments in TPR paradigms do not take these into consideration. Additionally, the awareness that one's behaviour and translation products will undergo particular scrutiny can further affect the translator's choices. Hence, the ecological validity of TPR results often suffers from the experimental setting of the studies, where translators will work in an unfamiliar environment and under irregular restraints. These caveats call into question the meaningfulness of conclusions about translation and the translation process based on target texts produced in such artificial settings, which is furthermore reinforced by the differences found between experimental and corpus-based results (for instance Heilmann et al. 2021; cf. Neumann et al. 2022, 123-124). They let corpus research appear to be superior in ecological validity despite the fact that translation corpora also don't necessarily adequately capture the translation process, since the published texts contained in corpora underwent editing (Bisiada 2018, Kruger 2017) so that they do not reflect the translator's choices alone.

In order to test the comparability of translations produced in experiments and at translators' usual workplace, we compiled a small corpus of elicited translations. 20 translators were commissioned to translate short English texts into German according to their usual workflow, in their usual working environment and at their own pace with access to resources of their preference. These English source texts had already been used in two previous translation experiments (Heilmann et al. 2021; Heilmann et al. 2022), which means that we had already collected German translations of these texts produced in our translation lab. In total, the corpus collected for this study consists of 220 texts (16432 words). Commissioning translators to translate the same text from home allows us to directly compare the translation choices made by each group and test in how far they relate to corpus findings.

The English texts were specifically designed for the lab experiments and represent short news-ticker style texts including a headline and three to five sentences. One set of source texts previously used in Heilmann et al. (2021) include a sentence containing a non-agentive construction that served as a controlled stimulus in our experiment. A non-agentive construction is the combination of an inanimate subject, such as *study* or *experiment*, with a verb that requires an agent, such as *show* or *find*. While non-agentive constructions exist in both languages, they are less common in German because the semantic mapping onto subjects is

more restricted in German (König and Gast 2018). Non-agentive constructions should, thus, pose a challenge to the translators given this contrastive difference. The source texts used in Heilmann et al. (2022) are centred around *of*-NP constructions (for example *the brain of the apes*). For this experiment, each text included two different types of NPs (Schönthal 2016), one *of*-NP of possession, for example *the engine of the spacecraft*, and one *of*-NP of engagement, which expresses a nominalised activity, for example *the departure of the spacecraft*. Heilmann et al. (2022) showed in a preceding corpus analysis that possession *of*-NPs are twice as frequent in English as engagement types, and they thus assumed that a more frequent grammatical construction would facilitate the translation process and would show more routinised and less diverse translation strategies in the translation product.

For this study, the translators were commissioned to translate the same texts, but this time from the comfort of their home, where they were free to use any kind of translation resource. They were also asked to provide information on the tools and resources they used, whether they used machine translation to pre-translate the source texts, and the time and breaks they took to finish the translations. These translations were then compared to the texts we collected in the lab experiments and to texts obtained from the translation corpus CroCo (Hansen-Schirra et al. 2012) to analyze how the difference in ecological validity affects the translation product. The texts were compared with regard to the translation of the two linguistic features, non-agentive constructions and *of*-NPs as well as more general linguistic characteristics such as average sentence length, metaphorisation and lexical density. We tested the results statistically with the help of linear mixed regression modelling. Preliminary results suggest that the translations produced in the lab-external condition have shorter sentences and are less explicit than the lab-internally produced texts.

These results put into relief the effect of the experimental setting on the translation product and illuminate the differences between experimental data and corpus data. With this study, we support the call for multi-methods approaches to overcome the limited ecological validity of lab experiments while still maintaining sufficient control over the setting.

References

Bisiada, Mario (2018). “The Editor’s Invisibility: Analysing Editorial Intervention in Translation.” In: *Target*, 30.2, pp. 288-309. https://doi.org/10.1075/target.16116.bis

Hansen-Schirra, Silvia, Stella Neumann, and Erich Steiner (2012). Cross-Linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German. Berlin: de Gruyter.

Heilmann, Arndt, Tatiana Serbina, Daniel Couto-Vale, and Stella Neumann (2019). “Shorter Than a Text, Longer Than a Sentence: Source Text Length for Ecologically Valid Translation Experiments”. In: *Target* 1, pp. 98–125.

Heilmann, Arndt, Tatiana Serbina, Jonas Freiwald, and Stella Neumann (2021). “Animacy and Agentivity of Subject Themes in English-German Translation”. In: *Lingua* 261 (October), 102813. https://doi.org/10.1016/j.lingua.2020.102813

Heilmann, Arndt, Jonas Freiwald, Stella Neumann, and Zoë Miljanović (2022). “Analysing the effects of entrenched grammatical constructions on translation.” In: *Translation, Cognition & Behavior* 5.1, pp. 110-143.

König, Ekkehard and Volker Gast (2018). *Understanding English-German Contrasts* (4th ed.). Erich Schmidt Verlag.

Kruger, Haidee (2017). “The Effects of Editorial Intervention: Implications for Studies of the Features of Translated Language.” In: *Empirical Translation Studies*. Ed. by Gert de Sutter, Marie-Aude Lefer and Isabelle Delaere. Berlin: de Gruyter, pp. 113-155.

Neumann, Stella, Jonas Freiwald, and Arndt Heilmann (2022). “On the use of multiple methods in empirical translation studies: A combined corpus and experimental analysis of subject identifiability in English and German.” In: *Extending the Scope of Corpus-based Translation Studies*, Ed. by Sylviane Granger and Marie-Aude Lefer. London: Bloomsbury, pp. 98-129.

Schaeffer, Moritz and Michael Carl (2017). “Language Processing and Translation”. In: *Empirical Modelling of Translation and Interpreting*. Ed. by Silvia Hansen-Schirra, Oliver Czulo, and Sascha Hofmann. Berlin: Language Science Press, pp. 117–154.

Schönthal, David (2016). “On the Multifaceted Nature of English Of-NPs: A Theoretical, Corpus, Cotextual and Cognitive Approach.” PhD Thesis, Cardiff University.

Exploring non-terminological formulaic sequences with corpora: closed-class keywords in translated law

Justyna Giczela-Pastwa

University of Gdańsk, Poland

KEYWORDS: legal translation, formulaicity, L2 translation, interference, translationese

Corpora have been successfully used in a number of research projects that focus on identifying differences between translated and non-translated language. Since 1993, when in her seminal paper Mona Baker called for the identification of 'universal features of translation (…) which are not the result of interference from specific language systems' (Baker, 1993, p. 243), translationese in several national languages has been thoroughly studied (cf. e.g. Alasmri & Kruger, 2018 for Arabic; Baroni & Bernardini, 2006 for Italian; Biel, 2014 for Polish; Cappelle & Loock, 2013 for French; Dayrell, 2004 for Brazilian Portuguese; Delaere et al., 2012 for Belgian Dutch; Labrador, 2011 for Spanish; Puurtinen, 2003 for Finnish; Saldanha, 2004 for English; Vintar, 2016 for Slovene; Xiao, 2010 for Chinese). The three decades brought to the fore not only the perspective of several national languages, but also a number of methodological proposals. The third code of translation (Frawley, 1984) has been studied through comparable corpora, both monolingual and multilingual, parallel corpora, and comparallel corpora (Gaspari, 2015). These various perspectives and methods have been brought together by a shared purpose: to gain a deeper insight into the nature of translated language, and thus better control the quality of translations.

In my paper, I intend to present a research design and findings of a corpus-driven study of Polish law translated into English. The analysis was aimed at pinpointing the discrepancies between translated and non-translated statutory English, and at identifying possible reasons for the observed divergence. The study was conducted with the use of a self-compiled study corpus that consisted of 30 Polish statutes translated into English (1.5m words), and three reference corpora of British, American and EU legislation (2.5-3.3m words). Every possible effort was made to ensure that the study corpus is balanced and representative: the examined statutes regulate a wide range of legal issues, and have been published by three different publishers. Nevertheless, the quantitative investigation was followed by a more qualitatively-oriented scrutiny whenever strong individual variation at the text level was observed, in order to identify all the possible determining factors. In addition to the study corpus and three reference corpora, a corpus of the original Polish statutes was constructed and consulted, in line with the reasoning that observing the realisation of a feature in the underlying source text enables the detection of true difference between translations and non-translations (Evert & Neumann, 2017, p. 49). This level of analysis seemed particularly justified in the project, as Polish law is frequently translated into English as L2, due to the status of Polish as a language of limited diffusion. Therefore, while discussing features of translationese it is necessary to consider the extent of interference from the source language (presumably stronger whenever it is the translator's L1) and strive to discriminate between the impact of a particular source language and the impact of the translation process as such. The corpus analysis software of choice was Sketch Engine (Kilgarriff et al., 2014).

Owing to divergent legal systems, it was expected that the translated English differs from non-translated legal English predominantly in terms of terminology and specialised phraseology. Initial findings in the project show that the translations are characterised by untypical collocations, i.e. typical lexical units are used in divergent combinations and with dissimilar frequency (Giczela-Pastwa, 2019). What is more, the translated English displays non-terminological patterns (i.e. text-structuring patterns, grammatical patterns, lexical collocations, cf. Biel, 2014) that are absent from or less frequent in non-translated English. This, in fact, may significantly account for a limited textual fit of the translations. In order to identify such patterns, a keyword analysis focused on closed-class tokens was carried out, inspired by the observation that "closed-class keywords can form a valid and even preferable basis for empirical linguistic research into specialized discourses" (Groom, 2010, p. 59). A systematic review of concordance lines for

the determiners and prepositions that were identified as positive keywords proved particularly helpful in discerning untypical non-terminological patterns in translated language.

The findings of the study may assist professionals in taking research-informed translation decisions in order to increase the textual fit of translated law, and could be of use for training purposes.

References

Alasmri, I., & Kruger, H. (2018). Conjunctive markers in translation from English to Arabic: a corpus-based study. *Perspectives*, *26*(5), 767–788. doi:10.1080/0907676X.2018.1425463

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

Biel, Ł. (2014). Lost in the Eurofog: The textual fit of translated law. Peter Lang.

Cappelle, B., & Loock, R. (2013). Is there interference of usage constraints? A frequency study of existential *there is* and its French equivalent *il y a* in translated vs. non-translated texts. *Target*, *25*(2), 252–275. doi:10.1075/target.25.2.05cap

Dayrell, C. (2004). Towards a corpus-based research methodology for investigating lexical patterning in translated texts. *Language Matters*, *35*(1), 70–101. doi:10.1080/10228190408566205

Delaere, I., De Sutter, G., & Plevoets, K. (2012). Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target*, *24*(2), 203–224. doi:10.1075/target.24.2.01del

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In G. de Sutter, M-A. Lefer & I. Delaere (Eds.), *Empirical Translation Studies. New methodological and theoretical traditions* (pp. 47–80). De Gruyter.

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, linguistic and philosophical perspectives* (pp. 159–175). Associated University Press.

Giczela-Pastwa, J. (2019). Inverse legal translation: a corpus-driven study of multi-word units related to the structure of translated statutory provisions. In Ł. Biel, J. Engberg, M. R. Martín Ruano & V. Sosoni (Eds.), *Research methods in legal translation and interpreting: Crossing methodological boundaries* (pp. 48–65). Routledge.

Groom, N. (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 59–78). John Benjamins.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36.

Labrador, B. (2011). A corpus-based study of the use of Spanish demonstratives as translation equivalents of English demonstratives. *Perspectives*, *19*(1), 71–87. doi:10.1080/0907676X.2010.481047

Liu, K., & Afzaal, M. (2021). Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *PLoS ONE*, *16*(6), e0253454. doi:10.1371/journal.pone.0253454

Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, *18*(4), 389–406.

Saldanha, G. (2004). Accounting for the exception to the norm: Split infinities in translated English. *Language Matters*, *35*(1), 39–53. doi:10.1080/10228190408566203

Vintar, Š. (2016). A bird's eye view of lexical creativity in original vs. translated Slovene fiction. *Perspectives*, *24*(4), 591–605. doi:10.1080/0907676X.2015.1069860

Xiao, R. (2010). How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, *15*(1), 5–35. doi:10.1075/ijcl.15.1.01xia

Trying one's hand at students' difficulties in sign language translation: A pilot study on errors in an LSFB-to-French learner translation corpus

Alice Heylens

University of Namur (NaLTT/LSFB-LAB), Belgium & UCLouvain (ILC/CECL), Belgium

KEYWORDS: Sign language, LSFB, learner translation corpus, error analysis, translation difficulties

Translation and interpreting programmes in sign languages are young educational curricula. In French-speaking Belgium, for example, the BA and MA programmes were launched in 2014 and 2017, respectively. However, the translation difficulties inherent in sign language translation have received very little attention to date, leaving students and lecturers alike with few pedagogical materials to rely on. Yet, sign and spoken languages can be considered as distant languages, both culturally and linguistically (cf. Nida, 1945). On the one hand, anthropologists recognize the existence of Deaf culture and Identity (Delaporte, 2002). On the other hand, linguists have identified specific properties of sign languages which impact their grammatical and discursive organization (e.g. being articulated in space and being expressed through different body parts such as the face and the hands) (Pfau et al., 2012). Translation from a sign language to a spoken language can therefore be expected to give rise to specific translation difficulties.

One of the ways of studying attested areas of difficulty in translation is through computer-aided error analysis in learner translation corpora (Granger & Lefer, 2020; Kübler et al., 2022; Verplaetse, 2022). Empirical insights gained from error analysis can then help tailor remedial exercises and improve translation curricula (Kunilovskaya et al., 2022). Error analysis is helpful to retrieve common difficulties faced by several students when translating a specific subject field and source language (Bowker & Bennison, 2003; Granger & Lefer, 2020). For example, in an English-to-Catalan learner translation corpus, Espunya (2014) found that 7 out of the 10 most frequent errors were related to target language accuracy such as Lexical Imprecision, Grammar, Punctuation and Spelling.

The present talk reports on the results of a pilot study devoted to the language pair LSFB (French Belgian Sign Language) and French. It relies on an error-annotated parallel corpus including source videos in LSFB and their French translations produced by students. The corpus is currently being collected within the framework of the international *Multilingual Student Translation* (MUST) project (Granger & Lefer, 2020) and is the first to include a sign language. The aim of the MUST initiative is to collect a large multilingual corpus of translations produced by foreign languages learners and student translators, together with a rich set of metadata about source texts (e.g. genre, format, communicative purpose), translation tasks (e.g. students status, duration, tools and resources used) and students (e.g. age, main languages of instruction, translation experience).The source videos used with the students are taken from the LSFB Corpus (Meurant, 2015), which contains 80 hours of semi-directed interviews where deaf people discuss in pairs. The corpus is representative of several regional varieties of LSFB and of different genres (narratives, conversations, explanations, argumentations and descriptions) (Meurant et al., 2016). A third of the videos are annotated: each sign is given an IDGloss, i.e. a written label of the lemma corresponding to the sign token. In this pilot study, the source video is a 2min01-long video in which two LSFB-native women discuss their sign names (where their sign names came from, who had given them, what they think about them). This source video will be translated by 19 master's students in LSFB-French translation and interpreting at UCLouvain. The reason for including interpreting students (n = 16) is two-fold: (1) the cohort of translation students is too limited (n = 3) and (2) interpreting students hold a BA in translation and interpreting and are therefore all trained in translation. While translating the video into French, they will not have access to the video's metadata, such as sign annotations, but they will be allowed to use tools and resources such as dictionaries.

The error annotation system used for the analysis is the second version of the *Translation-oriented annotation system* (TAS) (Granger & Lefer, 2021). The system, which has been developed collaboratively within MUST, contains 60 tags to annotate errors but also positive features such as good translation

choices. The taxonomy is hierarchical (up to four levels) as it contains different levels of granularity. TAS contains 8 first-level error categories (Content, Culture, Translation brief, Grammar and Syntax, Lexis and Terminology, Discourse and Pragmatics, Mechanics, Register and Style) and four meta-tags (PLUS for good translation choices, Source Language Intrusion, Distortion, and Error repetition to mark all occurrences of the same error throughout a given translation). First-level categories are further divided into sub-categories (e.g. Content: Distortion, Omission, Addition, Indecision, Non-identification of source-text error/inconsistency), which allows for fine-grained insights into translation quality. Students translations, together with corresponding metadata, will be collected through the Hypal4MUST tool, then sentence-aligned and error-annotated with the help of TAS on the teacher interface of the tool.

The main aim of the pilot study is to gain preliminary insights into the types of errors students make when translating from LSFB to French and thereby identify the main difficulties they face. The error analysis will be enriched with some of the collected metadata, such as the tools students report to have used while translating (the tools available for the LSFB-French language pair are strikingly different from those typically available for pairs involving two spoken languages). More generally, the pilot study also aims to assess the feasibility of data collection in terms of the time needed for the task at hand and the suitability of the TAS taxonomy to analyze translation errors from a sign language, which will help inform future larger-scale data collection.

References

Bowker, L., & Bennison, P. (2003). Student Translation Archive. In S. Bernardini, F. Zanettin, & D. Stewart, *Corpora in translation education* (pp. 103–117). Routledge. https://doi.org/10.4324/9781315760162

Delaporte, Y. (2002). *Les sourds c'est comme ça* (Éditions de la Maison des sciences de l'homme). Ministère de la culture.

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation*, *48*(1), 33–43. https://doi.org/10.1007/s10579-013-9260-1

Granger, S., & Lefer, M.-A. (2020). The Multilingual Student Translation corpus: A resource for translation teaching and research. *Language Resources and Evaluation*, *54*(4), 1183–1199. https://doi.org/10.1007/s10579-020-09485-6

Granger, S., & Lefer, M.-A. (2021). *Translation-oriented Annotation System manual (Version 2.0)* (CECL Papers 3). Centre for English Corpus Linguistics/Université catholique de Louvain. https://cdn.uclouvain.be/groups/cms-editors-cecl/TAS-2.0_annotation_manual_2021_TRAD-FR_final.pdf

Kübler, N., Mestivier, A., & Pecman, M. (2022). Using comparable corpora for translating and post-editing complex noun phrases in specialized texts: Insights from English- to-French specialized translation. In S. Granger & M.-A. Lefer, *Extending the Scope of Corpus-Based Translation Studies* (pp. 237–266). Bloomsbury Academic. http:// dx.doi.org/10.5040/9781350143289.0005

Kunilovskaya, M., Ilyushchenya, T., Morgoun, N., & Mitkov, R. (2022). Source language difficulties in learner translation: Evidence from an error-annotated corpus. *Target*, *35*(1), 34–62. https://doi.org/10.1075/target.20189.kun

Meurant, L. (2015). Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB). LSFB-Lab, University of Namur. www.corpus-lsfb.be

Meurant, L., Gobert, M., & Cleve, A. (2016). *Modelling a Parallel Corpus of French and French Belgian Sign Language*. 10th Language Resources and Evaluation Conference (LREC), Portoroz, Slovenia.

Nida, E. (1945). Linguistics and Ethnology in Translation-Problems. *Word*, *1*(2), 194–208. https://doi.org/10.1080/00437956.1945.11659254

Pfau, R., Steinbach, M., & Woll, B. (Eds.). (2012). *Sign language: An international handbook*. De Gruyter Mouton.

Verplaetse, H. (2022). Translation quality in student specialized translation: The impact of corpus use. In S. Granger & M.-A. Lefer, *Extending the Scope of Corpus-Based Translation Studies* (pp. 209–236). Bloomsbury Academic. http:// dx.doi.org/10.5040/9781350143289.0005

Questions in Mandarin and French: Contrastive insights from a bilingual comparable corpus of contemporary fiction

Hung-Hsin Hsu

Université catholique de Louvain, Belgium, & National Chengchi University, Taiwan

KEYWORDS: question types, question structures, comparable corpus of fiction, Taiwan Mandarin, French

The topic of questions has received relatively scant attention in contrastive studies to date (Axelsson, 2020; Coveney, 2011; Curry & Chambers, 2017), especially when it comes to pairs that involve languages other than English. Despite the fact that the relative dominance of English is gradually shifting, English still plays a significant role in contrastive studies, particularly in corpus-based research (Hasselgård, 2020). The present paper aims to contribute to filling this gap by examining questions in a Mandarin-French comparable corpus, paying particular attention to question types and their respective structures.

A distinction is generally made in typology between three question types (Siemund, 2001), namely polar questions (also known as *yes-no* questions; asking for confirmation), disjunctive questions (which contain at least two options from which the addressee can choose from), and *wh*-questions (also known as constituent questions; they include an interrogative word and the addresser is seeking information). Table 1 provides examples of these three question types in the two languages investigated in this study.

Table 1. Question types in Mandarin and French

	Mandarin	**French**	**English translation**
Polar questions	(1) *Nǐ shì xuéshēng* ***ma****?*	(4) *Es-tu un étudiant?*	Are you a student?
Disjunctive questions	(2) *Nǐ hē chá háishì kāfēi* ***(ne)****?*	(5) ***Est-ce que*** *tu bois du thé ou du café?*	Do you drink tea or coffee?
Wh-questions	(3) *Nǐ chī shénme* ***(ne)****?*	(6) *Qu'****est-ce que*** *tu manges?*	What are you eating?

As regards question structures, typological research shows that different strategies can be used to form questions across languages (Siemund, 2001, p. 1010; Ultan, 1978). These can be phonetic (e.g. intonation), morphological (e.g. interrogative particles, interrogative words) or syntactic (e.g. the order of constituents). Mandarin uses intonation and interrogative particles, such as *ma* in (1) and *ne* in (2) and (3), whereas French can rely on constituents reordering (i.e. inversion) in addition to intonation and interrogative particle, such as *est-ce que* in (5) and (6). Interestingly, in Mandarin, the *ne* particle in disjunctive and *wh*-questions is optional (Chu, 1998), while the particle *ma* is said to be required in polar questions (Huang et al., 2009, p. 237). In French, inversion and the *est-ce que* particle are both possible in all question types, albeit mutually exclusive (Siemund, 2001, p. 1022). In addition, interrogative words in French can be found in-situ, i.e. in the base-generated position, as in *tu manges quoi?* ('you eat what?').

Preliminary findings, based on Mandarin and French web data derived from the zhTenTen17 and frTenTen17 corpora (Jakubíček et al., 2013), show that in the two languages polar questions are more frequent than the other question types, which is in line with the theory of social economics of questions (Levinson, 2012) and the results of studies devoted to English (Siemund, 2017; Stivers, 2010). An interesting cross-linguistic difference that emerges from the web data is that *wh*-questions are significantly more frequent in French than in Mandarin, possibly due to the high number of different structures available in French to express *wh*-questions, namely inversion, declarative forms (i.e. intonation), *est-ce que* particle, *wh*-in-situ, and fragments (Coveney, 2011; Guryev, 2017, p. 1). In addition to this cross-

linguistic difference in question types, the pilot study shows that inversion represents the dominant structure across all question types in French, while the use of the *est-ce que* particle is marginal. The Mandarin web data reveal that, the *ma* particle is quite often omitted in polar questions, which goes against the claims commonly found in the literature.

However interesting, these trends remain tentative in view of the fact that web corpora contain all kinds of registers, which can potentially jeopardize the cross-linguistic comparability of the findings. Therefore, in the present paper, I build on these preliminary results and examine the distribution of question types on the basis of a 1-million-word self-compiled bilingual comparable corpus of contemporary fictional texts (detective stories and fantasy). More specifically, I set out to test two hypotheses: (1) polar questions are the most frequent question type in the two languages and (2) *wh*-questions are more widespread in French than in Mandarin. The comparable corpus was uploaded to Sketch Engine (Kilgarriff et al., 2014) and questions in the two languages were extracted by the CQL query [word = "\?"], which makes it possible to retrieve all question marks (see Axelsson, (2020); Biber et al., (1999, p. 211) for a similar approach). The query returns more than 3,000 occurrences in each language (see Table 2).

Table 2. Question marks in the self-compiled bilingual comparable corpus

	Mandarin	**French**
Raw frequencies	3,211	3,573
Relative frequencies per 100,000 tokens	770	649
Corpus size (tokens)	417,049	550,807

Some occurrences, namely the ones that do not express an interrogative mood, were weeded out from this initial dataset. Examples of discarded items include phone greetings (e.g. *âllo?*), name calling (e.g. *Mariam?*) and interjections (e.g. *Ah?*). Questions were then coded for question types and syntactic structures. In my talk, I will present these corpus findings, providing more detailed information about the structures used in Mandarin and French to form the three question types.

References

Axelsson, K. (2020). Questions in English and Swedish fiction texts: A study based on parallel and comparable corpus data. *Languages in Contrast*, *20*(2), 235–262. https://doi.org/10.1075/lic.00017.axe

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Chu, C. C. (1998). *A Discourse Grammar of Mandarin Chinese*. Peter Lang. https://www.peterlang.com/document/1087506

Coveney, A. (2011). L'interrogation directe. *Travaux de linguistique*, *63*(2), 112–145. https://doi.org/10.3917/tl.063.0112

Curry, N., & Chambers, A. (2017). Questions in English and French Research Articles in Linguistics: A Corpus-Based Contrastive Analysis. *Corpus Pragmatics*, *1*(4), 327–350. https://doi.org/10.1007/s41701-017-0012-0

Guryev, A. (2017). *La forme des interrogatives dans le corpus suisse de SMS en français: Étude multidimensionnelle* [Université de la Sorbonne Nouvelle (France) & Université de Neuchâtel (Suisse)]. https://www.theses.fr/2017USPCA045

Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast*, *20*(2), 184–208. https://doi.org/10.1075/lic.00015.has

Huang, C.-T. J., Li, Y.-H. A., & Li, Y. (2009). *The Syntax of Chinese*. Cambridge University Press. https://doi.org/10.1017/CBO9781139166935

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). *The TenTen Corpus Family*. 125–127.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Levinson, S. C. (2012). Interrogative intimations: On a possible social economics of interrogatives. In J. P. de Ruiter (Ed.), *Questions: Formal, Functional and Interactional Perspectives* (pp. 11–32). Cambridge University Press. https://doi.org/10.1017/CBO9781139045414.003

Siemund, P. (2001). Interrogative Constructions. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher and Wolfgang Raible (Ed.), *Language typology and language universals* (pp. 1010–1028). Walter de Gruyter.

Siemund, P. (2017). Interrogative clauses in English and the social economics of questions. *Journal of Pragmatics*, *119*, 15–32. https://doi.org/10.1016/j.pragma.2017.07.010

Stivers, T. (2010). An overview of the question–response system in American English conversation. *Journal of Pragmatics*, *42*(10), 2772–2781. https://doi.org/10.1016/j.pragma.2010.04.011

Ultan, R. (1978). Some General Characteristics of Interrogative Systems. In J. H. Greenberg (Ed.), *Universal of human language*. CA: Stanford University Press.

Looking under the hood: which linguistic features contribute to the source language classification of direct and indirect translations into Finnish?

Ilmari Ivaska[1] and Laura Ivaska[1,2]

[1]University of Turku, Finland, [2]Finnish Literature Society – SKS

KEYWORDS: cross-linguistic influences, indirect translation, literary translation, source language identification, supervised machine learning

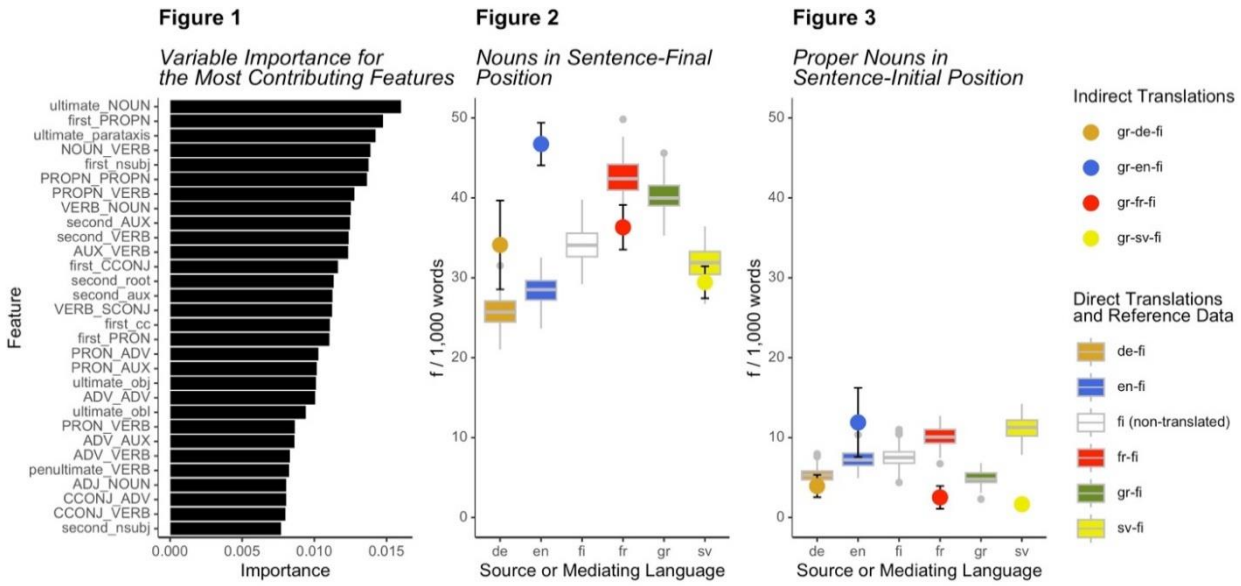
Automatic classification of translations according to their source languages (SLs) is a prominent line of research within corpus-based translation studies (Koppel & Ordan 2011; Lynch & Vogel 2012; Islam & Hoenen 2013; Volansky, Ordan & Wintner 2015). Such research has, on the one hand, been connected to language-pair-specific cross-linguistic influences, e.g. to see how the genealogical or typological closeness of source and target languages affects the classification (Rabinovich, Ordan & Wintner 2017). On the other hand, there has been discussion regarding the nature of linguistic phenomena that distinguish translated from non-translated language irrespective of the languages involved (e.g. Baroni & Bernardini 2006; Rabinovich & Wintner 2015). More recently, this line of research has been extended to indirect translations (ITr), that is, translations from translations. Research has shown that the language in which the original text was written contributes to the linguistic make-up of the final translation more than the language of the mediating text or any language-agnostic property related to translating (Rabinovich, Ordan & Wintner 2017; Ivaska & Ivaska 2022). However, research has so far paid little attention to the linguistic phenomena underlying this tendency. Building on the study by Ivaska & Ivaska (2022), this study addresses this research gap. While Ivaska & Ivaska, too, focused on the textual level to classify direct and ITrs according to their SL, here we focus on the linguistic nature of the mechanisms that underlie the successful classification of direct translations and the related unsuccessful classification of ITrs. The research questions are:

1) Which linguistic features contribute the most to identifying the SL of direct translations from various languages into Finnish?
2) How are these linguistic features distributed in ITrs into Finnish when the languages studied in research question 1 act as mediating languages, and the ultimate SL is kept constant?

Data and Method

The data are monolingual Finnish prose texts, and they include 1) direct translations from five SLs (English [en], French [fr], German [de], Greek [gr], and Swedish [sv]), 2) ITrs from gr as the ultimate SL and en, fr, de, or sv as the mediating language, and 3) a reference dataset of non-translated Finnish prose texts (altogether 137 texts and 8,467,580 tokens). The majority of the data stem from existing corpora but they are complemented with self-collected materials (for a full list, see Ivaska & Ivaska 2022: Appendices 1-3). All the data were parsed according to the Universal Dependencies annotation scheme (de Marneffe et al. 2021), using the Turku Neural parser (Kanerva et al. 2018).

The direct translations were divided into training and test data, so as to train and test a machine learning model to identify the SL of the translations based on the relative frequencies of a range of linguistic features, as well as to evaluate the importance that those features hold in the classification task. We use random forests as the machine learning algorithm (Breiman 2001), while the linguistic feature sets used include those that were shown to be most successful in SL identification of Finnish texts in an earlier



study (Ivaska & Ivaska 2022). Hence, the model includes as features sequential part-of-speech 2-grams (e.g., NOUN_VERB for a noun followed by a verb), positional 1-grams of syntactic dependencies (e.g., first_nsubj for a nominal subject in the sentence-initial position), as well as positional part-of-speech 1-grams (e.g., ultimate_NOUN for a noun in the sentence-final position) (altogether 363 features). These features are used to train a classifier and automatically classify the data into groups based on the SL. Then, the nature of the features that contribute the most to the classification is investigated in relation to the respective SLs: their distributional characteristics are explored in the direct translations, the ITrs and the reference dataset, so as to reveal the diverging characteristics of the various subgroups of data. Random forests provide a well-suited approach to the research problem at hand: due to the random sampling and inbuilt out-of-bag testing, they do not overfit easily. Also, most implementations of the algorithm, like the ranger package that we used (Wright & Ziegler 2017), also include a variable importance measure, which provides a means to evaluate the relative importance of different features.

Results and Discussion

The model reached an accuracy of 71.2% in predicting the SL of the test data (where baseline of the most common category, is 20%). The list of the thirty most important variables of the model (see Figure 1) indicates that the frequency of nouns in sentence-final positions (ultimate_NOUN) is the best predictor of the SL: as can be seen in Figure 2, sentence-final nouns (typically grammatical objects or obliques) are much less frequent in translations from de, en, and sv than in translations from fr and gr. In contrast, the second-best predictor, sentence-initial proper nouns (typically grammatical subjects), are less frequent in direct translations from gr than in all other data, as shown in Figure 3. Interestingly, in both cases three of the four datasets of ITrs diverge from the direct translations from the respective SLs towards the direct translations from gr. This can be interpreted to reflect cross-linguistic influences related to the dominant constituent orders of the respective languages: gr is the only SL in these data with no dominant constituent order, whereas en and sv have clear Subject–Verb–Object order, and de follows Subject–Verb–Object for main clauses and Subject–Object–Verb for dependent clauses (Dryer 2013).

In the presentation, we will explore the best distinguishing variables in detail in different language pairs, so as to reveal the typical differences in terms of the frequency distributions, and the lexical and functional variation underlying the differences. These preliminary results show that the method is suitable for revealing the differences that underlie successful SL identification. They also give an illustrative example on the nature of the linguistic phenomena that make ITrs more alike with direct translations from the respective ultimate SL rather than direct translations from the respective mediating language.

References

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259–274.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Dryer, M. S. (2013). Order of Subject, Object and Verb. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/81

Islam, Z., & Hoenen, A. (2013). Source and Translation Classification using Most Frequent Words. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 1299–1305. https://www.aclweb.org/anthology/I13-1185

Ivaska, I., & Ivaska, L. (2022). Source language classification of indirect translations. *Target. Special Issue: What Can Indirect Translation Research Do for Translation Studies?*, *34*(3), 370–394. https://doi.org/10.1075/target.00006.iva

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & Salakoski, T. (2018). Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 133–142. http://www.aclweb.org/anthology/K18-2013

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1318–1326. http://www.aclweb.org/anthology/P11-1132

Lynch, G., & Vogel, C. (2012). Towards the Automatic Detection of the Source Language of a Literary Translation. *Proceedings of COLING 2012: Posters*, 775–784. https://www.aclweb.org/anthology/C12-2076

Rabinovich, E., Ordan, N., & Wintner, S. (2017). Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 530–540. https://doi.org/10.18653/v1/P17-1049

Rabinovich, E., & Wintner, S. (2015). Unsupervised Identification of Translationese. *Transactions of the Association for Computational Linguistics*, *3*, 419–432. https://doi.org/10.1162/tacl_a_00148

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98–118. https://doi.org/10.1093/llc/fqt031

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Passive Constructions in En/Fr Human and Machine Translation

Daniel Henkel

University of Paris 8, France

KEYWORDS: Corpus linguistics, Translation, Machine Translation, Passive, Quantitative analysis

Introduction

Since it was first defined by Johansson in 2007, the "bidirectional" corpus model, combining both comparable and parallel corpora, has taken on increasing importance in CBTS research (cf. Frankenberg-Garcia 2009, Zanettin 2011, Vandevoorde et al. 2016, *inter alia*). The bidirectional approach allows for comparisons on multiple levels, which together provide more information and a more synthetic overview than studies based on comparable or parallel corpora alone:

- between original English (En0) and original French (Fr0), to observe differences unrelated to translation and establish benchmarks for each language,
- between subcorpora of target-texts and original texts in the same language, to determine whether target-texts follow or deviate from target-language norms,
- between pairs of source- and target-texts, to assess the amount of interlinguistic influence (a.k.a. "shining-through" Teich 2003).

Yet, as recently as 2022, Granger & Lefer regret the dearth of bidirectional studies. Admittedly, compiling bidirectional corpora is complicated, and resources are often limited by the availability of corresponding source and target-texts leading to trade-offs between comparability and representativity (Leech 2007).

The corpus and methodology used in the present study attempt to expand on the bidirectional approach in several ways:

- First, this corpus of 140 authors/translators is currently the largest and most diversified bidirectional, linguistically annotated and aligned corpus for English/French[1].
- Whereas corpus data are often expressed as totals or averages, the text-by-text approach is more informative as it reveals the amount of variation within each subcorpus.
- The integration of neural machine translations alongside human translations adds a new level of comparison between NMT and human translators
- Sentence-level alignment provides a means to overcome the limitations of macroscopic quantitative analysis through fine-grained qualitative analysis.

This study will seek to determine to what extent human or machine-produced target-texts deviate from target-language norms, the amount of influence from source-texts, and whether human and machine-produced translations can be distinguished as separate "subspecies". The indicator used as a basis for comparison will be analogous passive constructions in English (*be+past participle*) and French (*être+participe passé*).

2. Methods

2.1. Corpus preparation

[1]Other, larger bilingual corpora or corpus-like databases exist, of course, but which are not aligned or linguistically annotated. The closest contender, at the time of writing, is probably InterCorp.

Public domain literary works and corresponding translations from the late 19th century onward (median=1900) were collected from Project Gutenberg and Noslivres.net. A total of 35 works by 35 different authors in each language and the same number of translations (4×35 authors/translators = 140) were thus compiled into a ±13.5-million-word corpus consisting of four ±3.5m-word subcorpora. Source and target-texts were aligned automatically at sentence level with LF Aligner and double-proofread, once by student interns and then by the author. Machine translations were produced with DeepL Pro between late 2021 and early 2023, bringing the total to ±20m-words with the addition of two more ±3.5m-word subcorpora. Finally, all texts were tagged for POS and lemma in TreeTagger.

2.2. Data collection and analysis

Data were collected using regular expressions[2] in AntConc 3.5.9[3]. Results for each text were converted to normalized frequencies per 10,000 words (f/10k). Comparisons between subcorpora of target-texts and original texts representing the target-language were performed using the Wilcox-Mann-Whitney test, with Cohen's d as a measurement of effect-size. The correlation between source/target texts was evaluated using Spearman's[4] correlation coefficient. Statistical analyses were carried out in R.

3. Results

3.1. Original English vs. Original French

Passive constructions are more frequent in En0 (median 68.7/10k) than in Fr0 (median 41.4/10k):

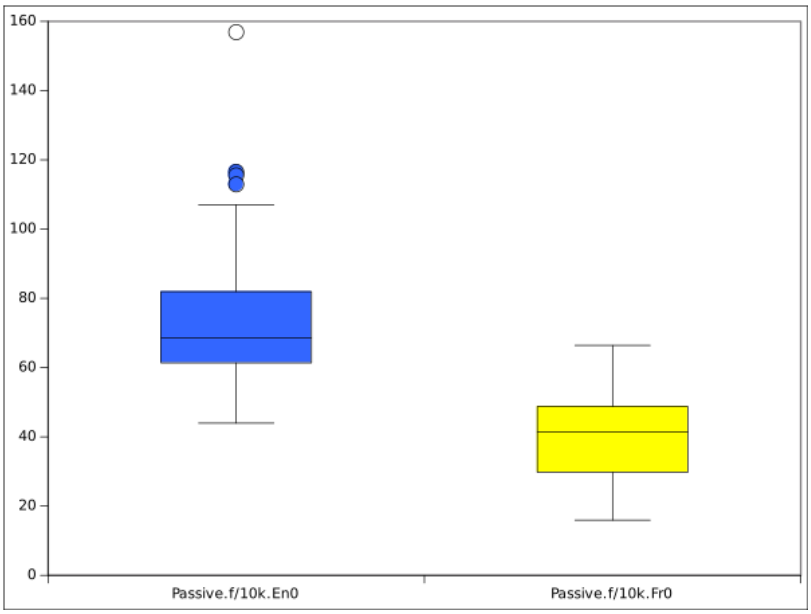


Figure 1. Passive constructions in Original English (En0) and Original French (Fr0): frequency per 10,000 words (f/10k)

3.2. Original English vs. EtrF

Nonetheless, the influence on target-texts is small in French-to-English translation (Fig. 2):

[2]Targeting passive constructions in English is relatively simple, whereas in French so-called unaccusative verbs such as *aller*, *tomber* as well as reflexive constructions such as '*il s'est levé*' must be inventoried as well and subtracted from the number of *être+participe passé* sequences.
[3]Unfortunately, in AntConc 4.x, due to limitations related to indexing, regular expressions are restricted to word-level and can no longer be used for complex syntactic queries at sentence-level.
[4]Spearman's *rho* was preferred so as to minimize the effect of outliers, but results with Pearson's *r* were statistically significant and largely similar.

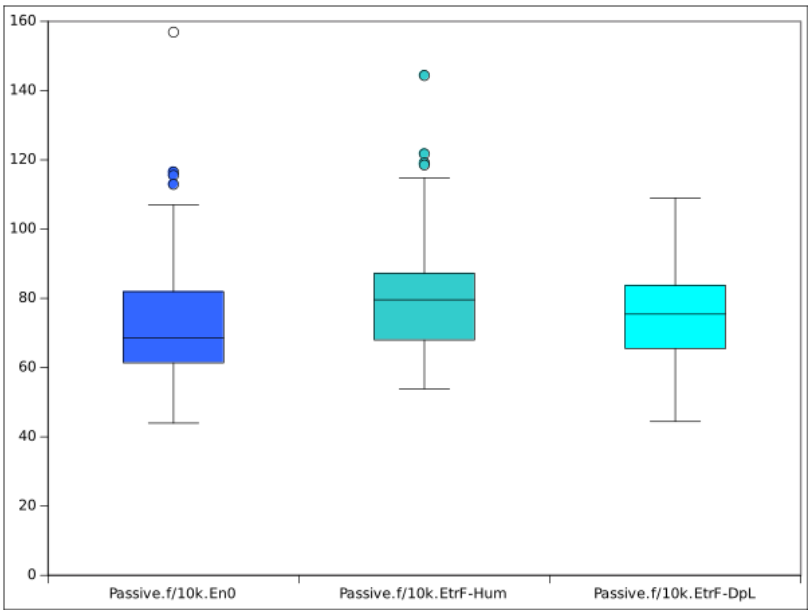


Figure 2. Passive constructions in Original English (En0), English-translated-from-French by Human translators (EtrF-Hum) and by DeepL (EtrF-DpL): frequency per 10,000 words (f/10k)

Contrary to expectations, human translators exhibit a small (d=0.35) but statistically significant (p=0.04) tendency to use the passive *more* than in En0 (+13.6%). The range of frequencies produced by DeepL is indistinguishable from En0 (p=0.49).

3.3. Original French vs. FtrE

In English-to-French translation (Fig. 3) the opposite occurs:

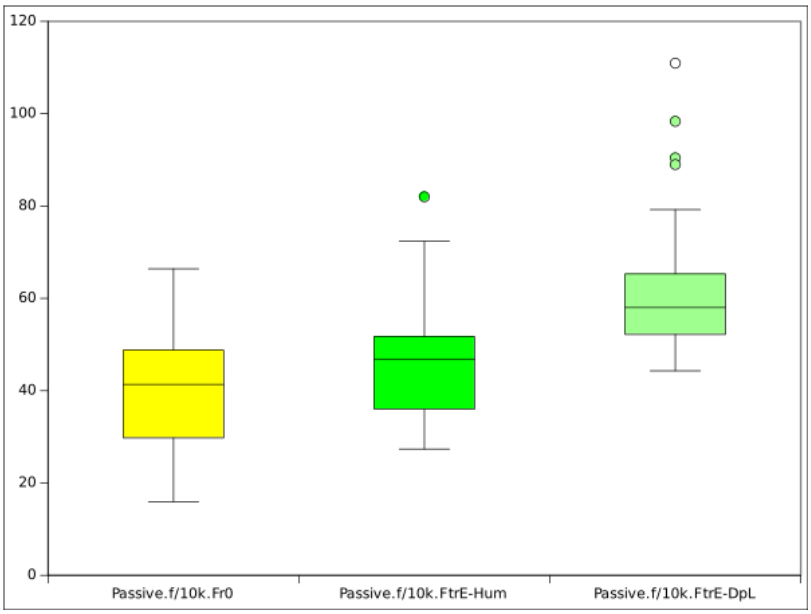


Figure 3. Passive constructions in Original French (Fr0), French-translated-from-English by human translators (FtrE-Hum) and by DeepL (FtrE-DpL): frequency per 10,000 words (f/10k)

Human translators use the passive more often in FtrE (d=0.42), but the difference is not statistically significant (p=0.14). In texts translated by DeepL, however, the passive occurs about 40% more often than in Fr0, a difference which is both quite large (d=1.55) and highly improbable by chance alone ($p<0.001$).

3.4. Source/Target-texts

Target-texts in FtrE produced by DeepL (Fig. 4b) are more strongly correlated (rho=0.94) to source-texts than those (Fig. 4a) produced by human translators (rho=0.72). In both cases the correlation is statistically significant (p<0.001):

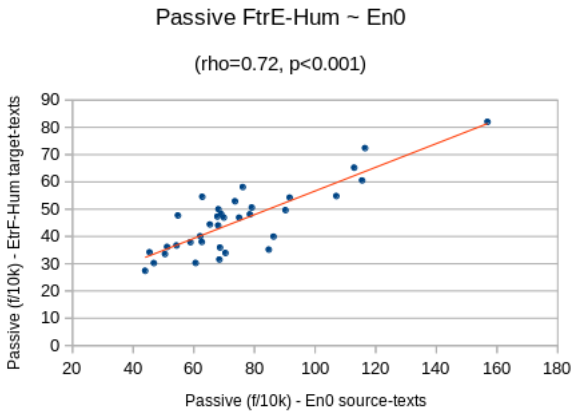


Figure 4a. Frequency of passives in corresponding source/target pairs: Human translators vs. En0

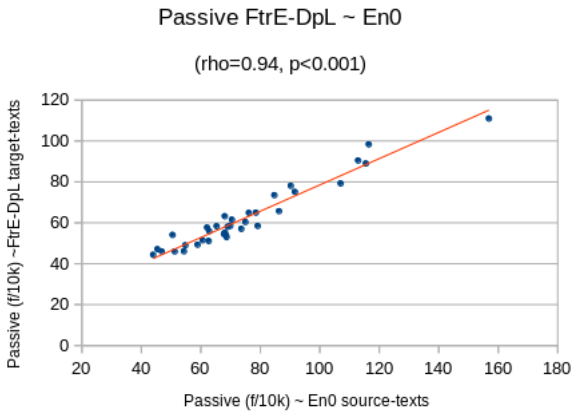


Figure 4b. Frequency of passives in corresponding source/target pairs: DeepL vs. En0

The correlation between source/target texts is statistically significant (p<0.001) for English-translated-from-French (Fig. 5a, 5b) as well, albeit somewhat weaker (rho=0.65 for EtrF-Hum, rho=0.87 for EtrF-DpL):

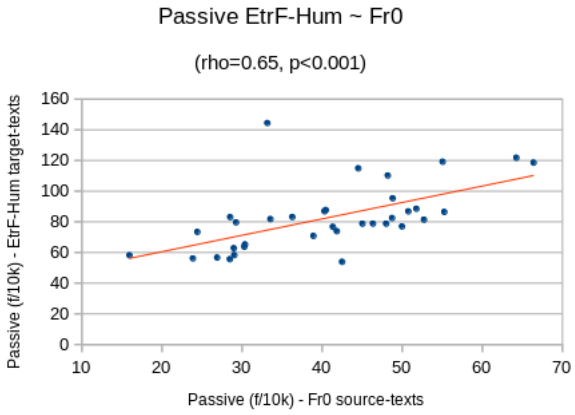


Figure 5a. Frequency of passives in corresponding source/target pairs: Human translators vs. Fr0

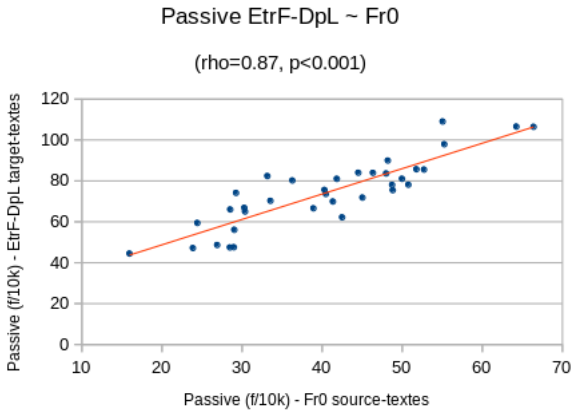


Figure 5b. Frequency of passives in corresponding source/target pairs: DeepL vs. Fr0

4. Discussion

Quantitatively, human translators' use of the passive in EtrF is different enough from En0 that EtrF can be recognized as a distinct subspecies. The same is true for machine translations in FtrE. These findings raise the following questions: Does DeepL outperform human translators in following target-language norms for English? Why does it fail to do so in the opposite direction?

Before these questions can be answered, it must be emphasized that quantitative data are only *suggestive* of underlying differences. If translators use the passive 13% more often in EtrF, it seems likely that they use it in different ways, but the nature of such differences can not be surmised from frequencies. Conversely, the lack of a quantitative difference between authors and translators or machine translation does not necessarily imply that they use the passive in exactly the same ways. Likewise, the stronger correlation coefficients suggest that DeepL is more imitative, but this remains to be confirmed through manual qualitative analysis.

One reason for the higher frequency of passives in EtrF is their use as a counterpart for the French pronoun '*on*'. This, however, is only one prominent example. Manual qualitative analysis of aligned source/target segments, which is currently underway, will yield further insight into when and where these two analogous constructions actually correspond to one another, what alternatives exist, and thus provide a comprehensive account of how human and machine translations differ in their use of the passive with respect to each other and to target-language norms.

References

Baker M. (1993). Corpus Linguistics and Translation Studies. Implications and Applications. In *Text and Technology*, M. Baker, G. Francis, G. and E. Tognini-Bonelli (eds.). Benjamins: Amsterdam & Philadelphia, 233–250.

Frankenberg-Garcia, A. (2009). Compiling and using a parallel corpus for research in translation. Babel: international journal of translation, 21(1), 57-71.

Granger, S. & M.-A. Lefer (2022). Corpus-based translation and interpreting studies: A forward-looking review. In *S. Granger & M.-A. Lefer (eds), Extending the Scope of Corpus-based Translation Studies*. Bloomsbury Advances in Translation series. London: Bloomsbury, 13-41.

Johansson, S. (2007). Seeing through multilingual corpora. In Corpus Linguistics 25 Years on. Brill Rodopi.

Leech, Geoffrey (2007) New Resources, or just Better Old ones? The Holy Grail of Representativness. In: *Hundt, Marianne/Nesselhauf, Nadja /Biewer, Carolin (eds) Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi, 133-149.

McEnery, T. and Xiao, R. (2007). Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist (Translating Europe)*. Multilingual Matters Ltd, Clevedon, UK. http://eprints.lancs.ac.uk/59/

Teich, E. 2003. Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts. Berlin: Mouton de Gruyter.

Vandevoorde, L., De Sutter, G., & Plevoets, K. (2016). On semantic differences between translated and non-translated Dutch: using bidirectional corpus data for measuring and visualizing distances between lexemes in the semantic field of inceptiveness. In *Empirical translation studies: Interdisciplinary methodologies explored*. Equinox Publishing, 128-146.

Zanettin, F. (2011). Translation and corpus design. SYNAPS – A Journal of Professional Communication 26/2011, 14-23.

Software

Anthony, L. (2020). AntConc (Version 3.5.9). Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Farkas, A. (2012). LF Aligner (Version 3.11 Linux). https://sourceforge.net/projects/aligner/

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Schmid, H. (1994-2021). TreeTagger, Universität Stuttgart, http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

The HANOI (Handwritten Notes of Interpreters) corpus for systematic study of note-taking in consecutive interpreting

Anna Jelec

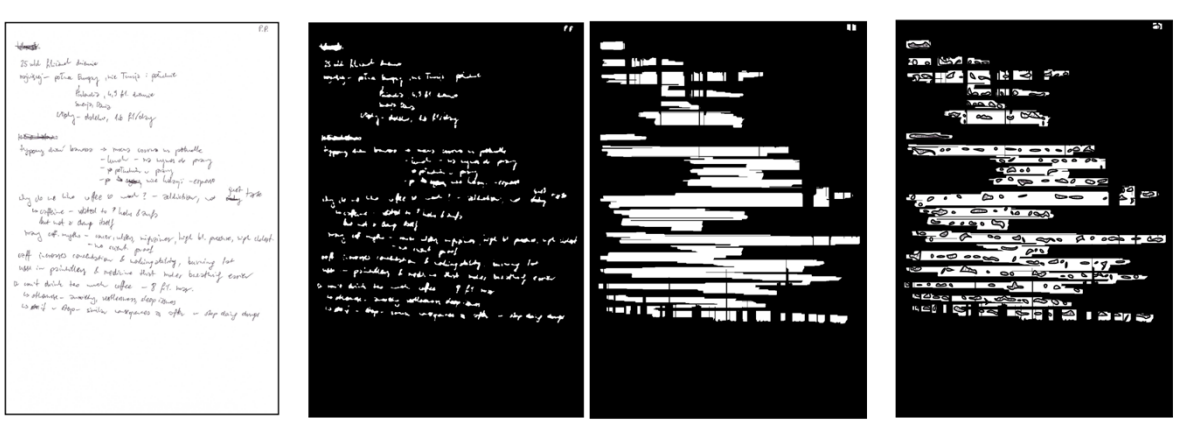
Adam Mickiewicz University, Poznań

KEYWORDS: note-taking, consecutive interpreting, interpreter training, online corpus, handwritten notes of interpreters

Notes in consecutive are a unique type of text. Each page is created at the intersection of two languages; the interpreter listens to the speech in the source language and notes down selected content in order to reproduce it in the target language. Although the terms *notes* and *note-taking* are relatively generic, here they refer to a specialised set of conventions and skills used in consecutive interpreting. Notes are a tool of the trade and a distinctive ability. In other words, the notebooks of two professional interpreters seldom look alike. This has raised certain questions for the didactics of interpreting, such as how to teach and assess note-taking skills or, indeed, whether to teach them at all (see, e.g., Chmiel 2010, Jelec 2015, Chmiel and Janikowski 2015, Bartłomiejczyk and Stachowiak-Szymczak 2022).

HANOI, or Handwritten Notation of Interpreters, is a corpus scanned notes collected from conference interpreters and interpreting students and a tool to support teaching note-taking in consecutive interpreting. As part of the DARIAH-PL Digital Research Infrastructure for the Arts and Humanities project, HANOI gathers material for large-scale systematic studies of note-taking for interpreting. It is a digital searchable database that allows users to search through scans of notes collected from working interpreters and students, carry out a visual analysis, search the corpus to find notes with specific parameters, download the corpus in its entirety and upload their own notes.

Figure 1. Stages in visual analysis: First, we scan the page, then we calculate page use, and finally from that we learn text density. The note had been made by an interpreting student prior to note-taking training



Step 1: the page is scanned, converted to grayscale and then to BW

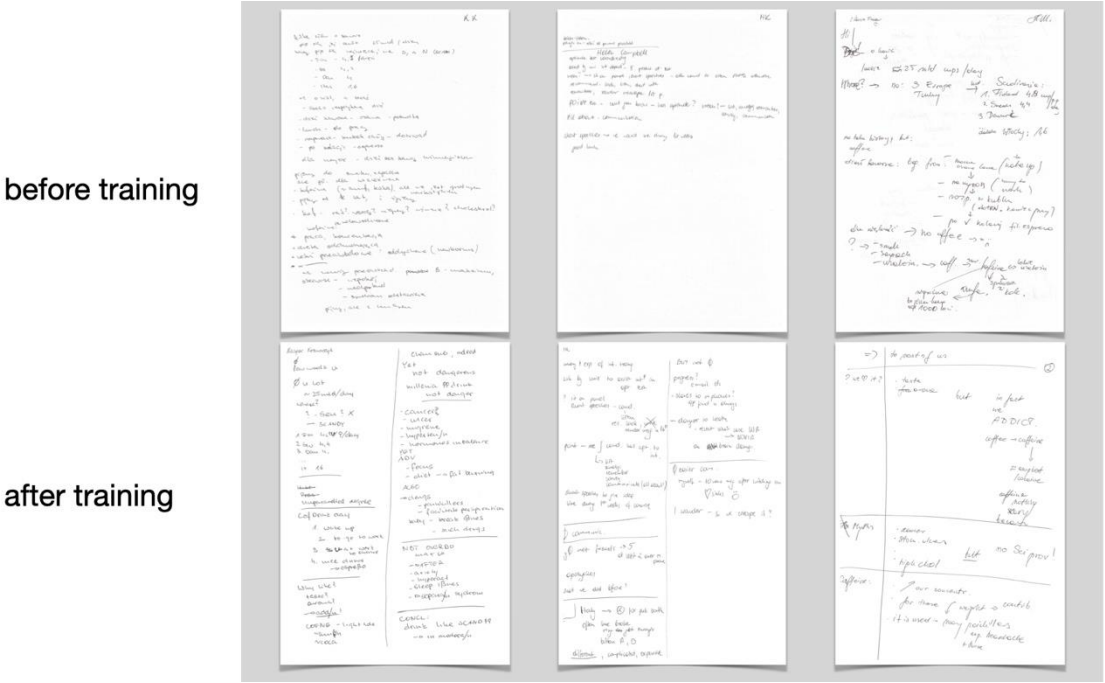
Step 2: page use is calculated as ratio of whole page to the area covered by pixel clusters (after padding is added to ink covered areas)

Step 3: text density is the ratio of page use vs. area covered in ink

The HANOI corpus allows researchers to ask about differences in the notes produced by interpreting students at various levels, and compare the notes of students and professionals. As shown in Figure 1, it currently uses two algorithms: one for the visual analysis of notes measuring the degree to which the interpreter writes information down (through their use of a blank page) and another for tracking the extent to which they structure notes (through text density). Using HANOI, we can study how the notes produced by interpreting students change during an introductory course in note-taking for consecutive (Figure 2, after Jelec et al. in press), or show how the notes of novice interpreters differ from those of seasoned professionals. By visualising the effects of teaching note-taking on students' notes, HANOI could support

the development of interpreter training curricula and assessment measures for teachers of consecutive interpreting, particularly in remote learning classrooms and student-led learning environments.

Figure 2. Examples of notes taken by students A, B and C before and after a course in note-taking. Student notes are part of the HANOI corpus.



Finally, HANOI allows users to upload their own notes to the corpus for automatic analysis. As a didactic tool, HANOI can help students identify a tendency to "over-note" (Sanchez 2018), which affects comprehension and attention (Gillies 2005), legibility, and delivery (Mead 2011).

At the moment, the corpus contains the scans of over 500 pages of notes, and is developed on an ongoing basis. The database of scanned notes of professionals is being actively expanded, allowing for more robust comparisons between the notes taken by interpreters at various stages of competence. Plans are also underway to include a tool for the automatic extraction of symbols in an effort to create a consecutive interpreters' symbol glossary.

References

Bartłomiejczyk and Stachowiak-Szymczak. 2022. Modes of conference interpreting: simultaneous and consecutive. In: Albl-Mikasa, Michaela and Elisabet Tiselius (Eds.). The Routledge handbook of conference interpreting. Taylor and Francis, 2022.

Chmiel, A. 2010. Interpreting Studies and psycholinguistics: A possible synergy effect. In: Daniel Gile, Gyde Hansen and Nike K. Pokorn (Eds.) Why Translation Studies Matters. Amsterdam: Benjamins, 223–236

Gillies, A. (2005). Note-taking for consecutive interpreting: A short course. St. Jerome.

Janikowski, P. (2015). Notowanie [Note-taking]. In P. Janikowski & A. Chmiel (Eds.), *Dydaktyka tłumaczenia ustnego* (pp. 141-166). Stowarzyszenie Inicjatyw Wydawniczych.

Jelec, A. (2015). Przetwarzanie w tłumaczeniu konsekutywnym [Processing in consecutive interpreting]. In: Chmiel, A., & Janikowski, P. (2015). *Dydaktyka tłumaczenia ustnego*. Stowarzyszenie Inicjatyw Wydawniczych.

Jelec, A., Stachowiak-Szymczak, K, Bartkiewicz, K and Ziobro-Strzępek, J. (in press) Why not(es)? Automatic Analysis of Notes for Consecutive Interpreting Training.

Mead, P. (2011). Co-ordinating delivery in consecutive interpreting. *inTRAlinea. on- line translation journal, 13*.

Pöchhacker, F. (2006). *Introducing interpreting studies* (Reprint). Routledge

Stachowiak-Szymczak, K. & Korpal, P. (2019). Interpreting Accuracy and Visual Processing of Numbers in Professional and Student Interpreters: an Eye-tracking Study. *Across Languages and Cultures*, 20: 2.

Sanchez, M. (2018). (Over)Note-taking in consecutive interpreting. *International Journal of English Language & Translation Studies*, *6*(3), 150–156.

EPICEL – A Simultaneous Interpreting corpus for Greek

Alina Karakanta

Leiden University, the Netherlands

KEYWORDS: corpora, interpreting, translation, Greek, European Parliament

Corpus-based Interpreting studies has been established as a blooming research field of translation, enriching the understanding of interpreted discourse. This type of research has been rendered possible through the availability and readiness of interpreting corpora, the most notable being the family of European Parliament Interpreting corpora, such as EPIC (Russo et al., 2005), EPICG (Defrancq et al., 2015), EPIC-UdS (Przybyl et al., 2022a) and PINC (Korzinek and Chmiel, 2021). These corpora allowed for comparative analyses of interpreted discourse vs original spoken discourse and interpreted vs translated texts (Kajzer-Wietrzny, 2012; Russo et al., 2012; Bernardini et al., 2016; Przybyl et al., 2022b). However, research in simultaneous interpreting (SI) has mainly focused on major languages such as English, Spanish, Italian, French and German. Even though Greek is an official language in the European Parliament (EP), to date there is no simultaneous interpreting corpus for the Greek language. In this work, we compile EPICEL, the first Greek SI corpus, from speeches of the EP. In this way, we contribute to the line of research in interpreted and, more generally, mediated discourse in the EP by focusing on Greek, an under-represented language in corpus-based interpreting studies which is less closely related to English in terms of vocabulary and structure.

The corpus contains i) original (EN) transcribed speeches, ii) their transcribed interpretations into Greek and iii) their written translations into Greek. To ensure consistency with the EP corpora family, the original speeches selected are the English talks from TIC (Kajzer-Wietrzny, 2012), which are then paired with their Greek interpretations. The Greek interpreting videos are first automatically transcribed and then edited, following the transcription guidelines of EPICG (Bernardini et al., 2018). The sentences are manually aligned to create a parallel EN→EL SI corpus. The comparable subcorpus containing the written translations of the respective speeches is aligned at the sentence level and allows for comparative research into the properties of translation vs interpreting.

We perform a preliminary analysis of the first 45 transcribed and sentence-aligned speeches using traditional corpus analysis measures. We compute standardised type-token ratio as a measure of lexical variation and mean sentence length. Our preliminary results are in line with previous research pointing out towards greater simplification in SI (Kajzer-Wietrzny, 2012; Bernardini et al., 2016; Przybyl et al., 2022b) and show that SI into Greek has lower lexical variation compared to written translations, as well as a shorter mean sentence length compared to both written translations and English spoken originals.

Further analyses will focus, among others, on the specificity of Greek as compared to other languages covered by TIC. Upon completion of the compilation process, sound-to-text alignment (Chmiel et al., 2022) and interpreter identification (Korzinek and Chmiel, 2021) will be added to facilitate further analyses into individual interpreter differences and prosodic features of simultaneous interpreting into Greek.

References

Bernardini, S., Ferraresi, A., Russo, M., Collard, C., & Defrancq, B. (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (Eds.), *Making way in corpus-based interpreting studies* (pp. 21–42). Springer, Singapore. doi: 10.1007/978-981-10-6199-8 2.

Bernardini, S., Ferraresi, A. & Miličević, M. (2016). From EPIC to EPTIC: Exploring simplification in interpreting and translation from an intermodal perspective. *Target, 28*(1), 61–86. DOI: 10.1075/target.28.1.03ber

Chmiel, A., Koržinek, D., Kajzer-Wietrzny, M., Janikowski, P., Jakubowski, D., & Polakowska, D. (2022). Fluency parameters in the Polish Interpreting Corpus (PINC). In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari

Ivaska & Silvia Bernardini (Eds.), *Mediated discourse at the European Parliament: Empirical investigations* (pp. 63–91). Berlin: Language Science Press. DOI: 10.5281/zenodo.697704

Defrancq, B., Plevoets, K., & Magnifico, C. (2015). Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to discourse and translation studies* (pp. 195–222). Springer, Cham. doi: 10.1007/978-3-319-17948-3 9.

Kajzer-Wietrzny, M. (2012). Interpreting universals and interpreting style. [PhD thesis]. Adam Mickiewicz University, Poznan.

Kajzer-Wietrzny, M., Ferraresi, A., Ivaska, I & Bernardini, S. (eds.). (2022). Mediated discourse at the European Parliament: Empirical Investigations (Translation and Multilingual Natural Language Processing 19). Berlin: Language Science Press.

Danijel Koržinek and Agnieszka Chmiel. (2021). Interpreter identification in the Polish Interpreting Corpus. *Revista Tradumàtica*, (19), 276–288.

Przybyl, H., Lapshinova-Koltunski, E., Menzel, K., Fischer, S., & Teich, E. (2022a). EPIC UdS-Creation and applications of a simultaneous interpreting corpus. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 1193–1200. European Language Resources Association (ELRA).

Przybyl, H., Karakanta, A., Menzel, K. & Teich, E. (2022b). Exploring linguistic variation in mediated discourse: Translation vs. interpreting. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (Eds.), *Mediated discourse at the European Parliament: Empirical investigations* (pp. 191–218). Berlin: Language Science Press. DOI: 10.5281/zenodo.697705

Russo, M., Bendazzoli, C., Sandrelli, A. & Spinolo, N. (2012). The European Parliament Interpreting Corpus (EPIC): Implementation and developments. In Francesco Straniero Sergio & Caterina Falbo (Eds.), *Breaking ground in corpus-based Interpreting Studies* (pp. 53–90).

Russo, M., Bendazzoli, C., Monti, C., Sandrelli, A., Baroni, M., Bernardini, S., Mack, G., Piccioni, L., Zanchetta, E., Ballardini, E., & Mead, P. (2005). *European parliament interpreting corpus*. http://catalog.elra.info/product_info.php?products_id=1145

Functional and cognitive profile of omissions in simultaneous interpreting

Eva Klüber and Kerstin Kunz

Heidelberg University, Germany

KEYWORDS: simultaneous interpreting corpus, cognitive load, omissions, semantic transfer, metafunctions of language

This paper presents an approach to identify omissions in an English-German parallel corpus of simultaneous interpretations and to characterize them in terms of several factors that increase or decrease the likelihood of information being omitted. Relating aspects such as types of information in the source text (ST) and different cognitive requirements for interpreting with the omissions produced we establish a functional as well as cognitive profile. Our talk will report on first findings which result from this approach.

Early studies into omissions hypothesize about the reasons for different types such as incomprehension or time management, and weigh their negative consequences on the target text (TT) (Barik, 1971). More recent approaches have partially reframed omission as a decision-based strategy, even rendering the speech more concise and coherent (Pym, 2009; Viaggio, 2002; Visson, 2005). Current research provides selective evidence into reasons and consequences of omissions, e.g. suggesting that they frequently follow explicitating shifts as an effect of cognitive load increase (Gumul, 2021). They further tend to occur in earlier sections of speeches and to contain both interpersonal elements and informative elements (see e.g. Dayter (2021) on the SIREN corpus). With our work, we hope to complement existing research by presenting a more comprehensive view. While omissions are often regarded as unintentional mistakes or conscious strategic decisions, we take a more gradual stance by relating each omission to the type of information omitted as well as the type of cognitive effort particularly contributing to overall cognitive load.

The corpus explored is part of HeiCIC (Heidelberg Conference Interpreting Corpus) and is a bidirectional parallel corpus of English and German originals and their simultaneous interpretations on the topic of automotive engineering (currently 117h, or 636,400 tokens). It is very homogenous in terms of information density, level of technicality, presentation and ST delivery speed and contains several interpretations for one original produced by student and professional interpreters, which allows us to draw a comparison between both groups. The corpus is transcribed and annotated on several layers using a semi-automatic process. We align ST and TT segments, identified as functional interpreting units, below the sentence boundary. TT segments are annotated for types of information transfer based on a semantic-pragmatic comparison of TT and ST segments. Main categories include addition, explicitation, implicitation and omission (Kunz et al., 2021).

We here focus on omission, defined as a segment whose (partial) content is present in the ST but not in the TT. It differs from implicitation, in particular, in that the ST content is not only not rendered in the TT, but also not recoverable from the co-text (Dayter, 2021; Gumul, 2017). In a next step, we establish a functional and cognitive profile of omission:

First, we use Halliday's metafunctions of language (Halliday, 2004) to distinguish types of information, characterizing segments of the ST (Gumul, 2021, 2022) as textual, ideational, or interpersonal. This provides insights into the importance individual pieces of information have for conveying the ST message and allows us to model variation in omissions in terms of their functional effect on the TT message.

Omission is further related to types of cognitive load as influencing factor. We base our study on cognitive load models that account for simultaneity of different cognitive efforts as well as for fluctuating requirements across ST and TT production (Gile, 2009; Seeber, 2007). We particularly focus on ST comprehension, short time memory and TT production as requirements/efforts determining overall cognitive load.

Frequencies of textual problem triggers in the ST (Mankauskienė, 2016) are used to identify increased cognitive load for ST comprehension. Semi-automatic annotation captures different subtypes, e.g. terminology, collocations, complex phrases, numbers and named entities. Occurrences of problem triggers are furthermore assessed in terms of their information status per text. First mentions typically introduce new information and should involve higher cognitive load than repetitions.

Memory requirements are captured by measuring ear voice span (EVS) (the time lag between the utterance of an ST segment and the aligned TT segment), assuming that EVS increases with elements stored in working memory. We automatically calculate the average EVS per interpreted speech in milliseconds as well as EVS for each segment and consider segment EVS that surpasses average speech EVS (a positive delta between both values) as a potential indicator of increased memory load.

We assess the five types of transfer categories introduced above in terms of cognitive requirements for their production on a scale, depending on their structural and semantic complexity in order to account for fluctuating demand in TT production effort. We assume equivalent transfer as the baseline of production load and compare deviating solutions, such as implicitation, or addition in their required load to give an indication of increased or decreased production load. For instance, addition may involve adding linguistic structures on top of the expression already present in the ST and may thus require increased production effort compared to a structurally equivalent solution. Omission should pose the lowest cognitive requirements for target text production in comparison to all other categories.

The criteria listed above are brought together for each text segment, attributing it with its functional and cognitive profile. Our approach thus allows us to correlate omissions with varying indicators of cognitive load increase or decrease and information type. As a consequence, we are able to present segment profiles with their likelihood to be omitted. In order to consider spill-over of cognitive load (Plevoets & Defrancq, 2021) we also take into account the previous segment's cognitive processing requirements. In addition, our corpus design enables us to report omissions by participant groups by separating students and professional interpreter data. The comparison of the groups indicates how the application of omissions may develop with increasing experience.

Results of a limited case study have indicated that omissions are often motivated by increased cognitive processing load of preceding segments. In particular, an excessively increased EVS in the previous segment correlates with omissions. The majority of observed omissions were omissions of entire segments rather than partial omissions, which may contribute more to cognitive relief. In addition, omitted segments were shorter, contained on average fewer PTs and served less often an ideational function than average segments.

We hope to confirm the preliminary results in a larger data set, as presented, and test our results for statistical significance. We will moreover relate omissions to language direction and directionality and attempt to determine the breadth of variation between several interpretations of the same original, also in different stages of experience. In further studies, we plan to extend the analysis to other categories of information transfer to gain further insights into the changing requirements of cognitive processing during simultaneous interpreting.

References

Barik, H. C. (1971). A Description of Various Types of Omissions, Additions and Errors of Translation Encountered in Simultaneous Interpretation. *Meta: Journal des traducteurs*, *16*(4). https://doi.org/10.7202/001972ar

Dayter, D. (2021). Strategies in a corpus of simultaneous interpreting. Effects of directionality, phraseological richness, and position in speech event. *Meta*, *65*(3), 594-617. https://doi.org/10.7202/1077405ar

Gile, D. (2009). *Basic concepts and models for interpreter and translator training* (Revised edition ed.). John Benjamins Publishing Company. http://ebookcentral.proquest.com/lib/ub-heidelberg/detail.action?docID=622263

Gumul, E. (2017). Explicitation and directionality in simultaneous interpreting. *Linguistica Silesiana*, *38*, 311-329.

Gumul, E. (2021). Explicitation and cognitive load in simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*, *23*(1), 45-75. https://doi.org/10.1075/intp.00051.gum

Gumul, E. (2022). Changing the text through explicitation. How trainee interpreters perceive the role of explicitating shifts. *Translator (Manchester, England)*, *28*, 17.

Halliday, M. A. K. (2004). *An introduction to functional grammar* (3. ed.). Arnold.

Kunz, K. A., Stoll, C., & Klüber, E. (2021). HeiCiC: A simultaneous interpreting corpus combining product and pre-process data. MOTRA,

Mankauskienė, D. (2016). Problem Trigger Classification and its Applications for Empirical Research. *Procedia - Social and Behavioral Sciences*, *231*, 143-148. https://doi.org/https://doi.org/10.1016/j.sbspro.2016.09.083

Plevoets, K., & Defrancq, B. (2021). Imported load in simultaneous interpreting. In R. Muñoz Martín & S. L. Halverson (Eds.), *Multilingual Mediated Communication and Cognition* (1 ed., pp. 18-43). Routledge. https://doi.org/10.4324/9780429323867-2

Pym, A. (2009). On omission in simultaneous interpreting: Risk analysis of a hidden effort. *Efforts and Models in Interpreting and Translation Research*, 83-105. https://doi.org/10.1075/btl.80.08pym

Seeber, K. G. (2007). Thinking outside the cube: Modeling language processing tasks in a multiple resource paradigm. In *Interspeech 2007* (pp. 1382-1385). https://archive-ouverte.unige.ch/unige:97484

Viaggio, S. (2002). The quest for optimal relevance: The need to equip students with a pragmatic compass. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century: Challenges and opportunities* (pp. 229-244).

Visson, L. (2005). Simultaneous Interpretation: Language and Cultural Difference. In S. Bermann & M. Wood (Eds.), *Nation, Language, and the Ethics of Translation* (pp. 51-64). Princeton University Press.

Comprehensibility of health information– comparing the use of direct and indirect address in standard German and Plain German health information

Janina Kröger

University of Hildesheim, Germany

KEYWORDS: plain language, health information, accessible communication, comprehensibility, intralingual translation

This research is funded by the Robert-Bosch-Stiftung via the interdisciplinary stipend programme "Chronic Diseases and Health Literacy"("*Chronische Erkrankung und Gesundheitskomeptenz*").

Introduction

Health information has to inform patients extensively about diseases, different therapies, and risks to enable patients to make an informed decision about their health (Lühnen et al. 2017). Health information place enormous demands on their recipients, as they convey complex specialist knowledge using the variety of linguistic means accessible in the language. However, a low health literacy hinders successful comprehension of health information. Health Literacy "entails people's knowledge, motivation and competences to access, understand, appraise and apply health information" (Sørensen et al. 2012, 3). Studies have shown that more than 50% of the German population has a low health literacy (Schaeffer et al. 2016, 2020) and cannot use information for their own health. Especially vulnerable target groups, e.g., people with communication impairments, people with low socioeconomic background or the elderly, are confronted with different barriers when interacting with health information (Rink & Maaß 2022).

Intralingual translation (Jacobson 1959) can be defined as overcoming barriers in communication (Hansen-Schirra & Maaß 2019, 3) and encompasses strategies such as translation into Easy Language or Plain Language to improve access to information (Rink & Maaß 2022). The National Action Plan Health Literacy recommends the use of Plain German to reach out to vulnerable target groups and improve comprehensibility of health information (Schaeffer et al. 2018, 43). "Plain Language is a dynamic variety that adapts to the needs of specific user groups in special target situations" (Maaß 2020, 150). While Easy Language has a fixed set of rules the translator has to follow, Plain Language encompasses a variety of characteristics. Plain Language in Germany can be conceptualised as a strategical enrichment of Easy Language rules, following the complexity of each grammatical feature (Bredel & Maaß 2016, Maaß 2020, 155ff) and the competences of the target group. One feature of Easy Language is directly addressing the reader whenever in line with the text function (Maaß 2020, 128). This improves the explicitness and therefore comprehensibility as well as the action-enabling potential of the texts (ibid).

There is no research yet as to which barriers are present in health information, and which strategies are used in intralingual translation to reduce these barriers. The presentation outlines the underlying research project while focusing on how direct and indirect address of the readers contribute to barriers within the texts.

Method

A corpus was compiled in June 2020. The corpus consists of twelve texts about chronic diseases published by the Wort & Bild publishing house on their website "apotheken-umschau.de". The "*Apotheken Umschau*" is an online platform for health information which also offers approximately two hundred texts in Plain German. The Plain German texts are a product of intralingual translation and linked directly to their standard German source texts. The topics of the texts in the analysed corpus are *arthrosis*, *asthma*, *Hashimoto's thyroiditis, Parkinson's disease, neurodermatitis*, and the drug *L-thyroxine*. The length of the texts varies between 985 and 5.021 words (standard language versions) and 558 to 948 words (Plain Language versions). A qualitative analysis was done using the software MaxQDA (Verbi GmbH 2022). Codes were developed that capture linguistic and conceptual complexity. The codes derive from features

of expert language (Roelcke 2010), e.g., terminology, nominal sentence structure, passive voice, and compound nouns, as well as from the texts themselves, (see Kröger & Ahrens 2022, Kröger in prep.). The coded segments were analysed with regards to their impact on comprehensibility. The Plain German texts are coded with the same codes. A comparison of the coded segments in standard German and Plain German was done to determine strategies used to improve comprehensibility in the Plain German texts.

Results

The presentation focuses on direct and indirect address in the texts. The code is chosen to illustrate how qualitative corpus analysis can help determine the complexity of texts. Preliminary results of the corpus analysis show that the standard German health information directly address the reader in 19 cases. The direct address has two main functions: They point the reader towards additional sources, e.g., the doctor or other online sources, and they prompt the reader to act a certain way, e.g., take the medicine at a specific time.

Impersonal words were tagged as a means for indirect address. Impersonal words included for example the use of the pronoun "*man*" ("*one"*) and "*es gibt*" ("*there is"*). Impersonal words are used in the source texts to state generic information, and to indirectly prompt the reader to act. Impersonal words are used more frequently than direct address which leads to the conclusion that the function of the standard German texts lies more on generic information than specific information on how to act in a certain case. As a result, the information is distanced from the specific reader. This may contribute to a cognitive barrier (Schubert 2016, Rink 2020). It also influences an emotion barrier (Lang 2021) as the distance in presentation allows the readers to distance themselves from the content. This can reduce emotional involvement with a potentially emotional subject, such as being personally affected by a disease.

The Plain Language texts are analysed in the same way. 90 segments are identified that directly address the reader. The analysis shows a shift in function, as the segments are mostly used to prompt the reader to act. The Plain German texts are more focused on giving information about how to act in a certain situation. At the same time, the use of impersonal words is reduced. 31 segments are using impersonal words. While the Plain German texts also give generic information, the presentation of knowledge about action is shifted between standard German and Plain German. While the standard German texts use impersonal words to also present knowledge about required action, Plain German texts shift to directly addressing the reader to convey how to act in a certain case.

References

Bredel, Ursula/Maaß, Christiane (2016): Leichte Sprache. Theoretische Grundlagen, Orientierung für die Praxis. Duden.

Hansen-Schirra, Silvia & Maaß, Christiane (2019): Translation proper – Kommunikationsbarrieren überwinden. University of Hildesheim: Hildok. https://doi.org/10.25528/015

Jakobson, Roman (1959): On Linguistic Aspects of Translation. In: Browner, Reuben Arthur (ed.): *On Translation* (pp. 232-239), Harvard University Press.

Kröger, Janina & Ahrens, Sarah (2022): An interdisciplinary approach to translation-oriented text analysis in intralingual translation research. Research Report. In: Ahrens, Sarah et. al (eds): *Accessibility – health Literacy – health information. Interdisciplinary Approaches to an Emerging Field of Communication.* (Easy – plain – accessible vol. 13, pp.113 – 138) Frank & Timme.

Kröger, Janina (in preparation): Verständlichkeit und Handlungsorientierung von Gesundheitsinformationen im Internet. Eine korpusgestützte Analyse am Beispiel "medizinische Ratgebertexte in Einfacher Sprache". [Manuscript of dissertation thesis]. University of Hildesheim.

Lang, Katrin (2021): Auffindbarkeit, Wahrnehmbarkeit, Akzeptabilität. Webseiten von Behörden in Leichter Sprache vor dem Hintergrund der rechtlichen Lage. (Easy – plain – accessible vol. 10) Frank & Timme.

Lühnen J, Albrecht M, Mühlhauser I &Steckelberg A. (2017) *Leitlinie evidenzbasierte Gesundheitsinformation.* http://www.leitlinie-gesundheitsinformation.de/.

Maaß, Christiane (2020): Easy Language – Plain Language – Easy Language Plus. Balancing Comprehensibility and Acceptability. (Easy – plain – accessible vol. 3.)Frank & Timme.

Rink, Isabel (2020): Rechtskommunikation und Barrierefreiheit. Zur Übersetzung juristischer Informations- und Interaktionstexte in Leichte Sprache. (Easy – plain – accessible vol. 1) Frank & Timme.

Rink, Isabel & Maaß, Christiane (2022): Verständlichkeit und Gesundheitskompetenz im Spektrum zwischen Leichter und Einfacher Sprache. In: Rathmann, K. et al. (eds.): *Gesundheitskompetenz*, Springer Reference Pflege – Therapie – Gesundheit, https://doi.org/10.1007/978-3-662-62800-3_146-1

Roelcke, Thorsten (2010): *Fachsprachen.* Schmidt.

Schaeffer, Doris, /Hurrelmann, Klaus, Bauer, Ullrich &Kolpatzik, Kai (2018): Nationaler Aktionsplan Gesundheitskompetenz. Die Gesundheitskompetenz in Deutschland stärken. KomPart.

Schaeffer, D., Vogt, D., Berens, E. M. &Hurrelmann, K. (2016): *Gesundheitskompetenz der Bevölkerung in Deutschland - Ergebnisbericht.* Universität Bielefeld.

Schaeffer, D., Berens, E., Gille, S., Griese, L., Klinger, J. ,de Sombre, S.,Vogt, D. & Hurrelmann, K. (2021): *Gesundheitskompetenz der Bevölkerung in Deutschland vor und während der Corona Pandemie. Ergebnisse des HLS-GER 2.*Interdisziplinäres Zentrum für Gesundheitskompetenzforschung (IZGK), Universität Bielefeld.

Schubert, Klaus (2016): Barriereabbau durch optimierte Kommunikationsmittel: Versuch einer Systematisierung. In: Mälzer, Nathalie (ed.): *Barrierefreie Kommunikation – Perspektiven aus Theorie und Praxis.* (pp. 15-33). Frank & Timme.

Sørensen, Kristine, van den Broucke, Stephan, Fullam, James, Doyle, Gerardine, Pelikan, Jürgen, Slonska, Zofia & Brand, Helmut(2012): Health literacy and public health: A systematic review and integration of definitions and models. *BMC Public Health 12(80).* DOI: 10.1186/1471-2458-12-80.

Analysis of automatically tagged discourse relations in translated and interpreted English

Ekaterina Lapshinova-Koltunski[1], Kaveri Anuranjana[2], Frances Pik Yu Yung[2], Christina Pollkläsener[1], Merel Scholman[2], Vera Demberg[2]

[1]University of Hildesheim, Germany, [2]Saarland University, Germany

KEYWORDS: discourse relations, discourse parser, translation, interpreting, spoken language

The present paper deals with the analysis of discourse connectives in English texts that were either originally produced in English (both written and spoken) or translated and interpreted. We explore the distributions of discourse connectives that trigger various relations in translated and non-translated language linking the observations to the phenomena of explicitation and implicitation (Klaudy and K aroly, 2005; Blum-Kulka, 1986). Relying on the existing studies on these phenomena in interpreting and translation, as well as spoken and written language (Lapshinova-Koltunski et al., 2022, 2021; Scholman et al., 2021; Crible, 2020; Crible and Cuenca, 2017; Kajzer-Wietrzny, 2012; Gumul, 2006; Shlesinger, 1995), we assume that in general, translated language contains more explicit discourse relations than the original one due to explicitation – a tendency to transfer the content of the source into the target in a more explicit way. However, (1) there is a difference in distribution of discourse connectives that trigger different relation types - as cognitively more complex relations (e.g. condition and comparison) cannot be easily left out, which means that they are expected to be more often expressed with discourse connectives than other relation types (see more details in Hoek et al., 2017). Moreover, (2) there are more discourse connectives in translation than in interpreting, as previous studies showed that the process of implicitation is frequent in interpreting. Apart from this, we expect to find source-language dependent differences. We also expect that the repository of discourse connectives to express discourse relations is reduced in spoken language, which means that we would find more variation in connectives used to express the same relation in written than spoken language.

For our analyses, we use the data from the European Parliament which includes officially published original speeches, the transcripts of the speeches delivered at the European Parliament, as well as their officially published translations and transcribed interpretations. The written data (published speeches and translations) originate from the corpus Europarl-UdS (Karakanta et al., 2018), while the transcribed data originates from the corpus EPIC-UdS (Przybyl et al., 2022). The translations under analysis were produced from German into English. The interpreted English has both German and Spanish as a source. All subcorpora are comparable, as they contain European Parliament speeches. Overall, we analysed over 10 million tokens of corpus data, with spoken data comprising a much smaller portion (ca. 170 thousand tokens).

We annotated discourse connectives with the help of the Discopy discourse parser (Knaebel, 2021; Knaebel and Stede, 2020), which is a BERT-based end-to-end discourse parser. We only utilized the explicit discourse relation classifier component. The parser classifies connective and non-connective usage of explicit connective tokens and assigns the type of discourse relation marked to the connective in each instance. The parser is trained on PDTB 2.0 (Prasad et al., 2008) to predict 14 second-level discourse relation senses. The accuracy to distinguish connective and non-connective usage, and to classify the sense of each connective are 97.18% and 86.26% respectively. In our analysis, we first focus on the distribution of the four upper categories: Expansion (e.g. and, also), Contingency (because, if, so), Comparison (but, however) and Temporal (finally, secondly). We also look into the subcategories of relations that are believed to be more complex from a cognitive perspectives and, therefore, may expose diverging tendencies in original and and translated, as well as spoken and written language. They include Contingency.Condition, Contingency.Cause, Comparison.Concession, Comparison.Contrast, Comparison.Similarity. We extract the distribution of the connectives according to the categories automatically assigned by the parser and compare them across the corpora under analysis. In general, we consider a higher number of connectives triggering a certain type of relation as a signal of explicitation.

In the following, we report on our observations on the distribution of various relations in original spoken and written texts, as well as interpreting and translation. The percentage of discourse relations is calculated against the total number of connectives identified by the parser in the corpus. We found that:

(1) Translated and interpreted texts are more explicit than comparable originals for certain relation types only. This holds for Expansion (46.87% in originals vs 47.95% in translations) and Contingency (23.48% in originals vs 25.99% in translation) in the written subcorpora, but only Contingency in the spoken language. Here, the number vary depending on the source language involved: 23.09% in original spoken English, 26.06% in English interpreted from German and 25.40% in English interpreted from Spanish.

(2) Interpreted English contains less Comparison (13.81% in German-English and 14.52% Spanish-English interpreting vs. 19.60% in English translated from German) and Temporal (4.87% and 4.84% in interpreting vs. 6.47% in translations) relations than translated English. However, it contains more Expansion (55.25% and 55.44% vs. 47.95%).

(3) In terms of Contingency relations, we found source-language-dependent differences: while English interpreted from German contains more Contingency.condition relations than originally spoken English (9.45% vs. 7.32%), English interpreted from Spanish contains less connectives labelled with this relation type (6.56%). An opposite trend is observed for Contingency.Cause: here, interpreting from Spanish contains more connectives than of this type (17.65%) than original English (14.40%) as well as interpreting from German (15.68%).

(4) Another interesting observation was made for Comparison.Contrast: while English interpreted from German contains less connectives of this type than originally spoken English, English translated from German contains more connectives of this type than originally written English.

(5) In terms of connectives, the top five most frequent connectives expressing Contingency overlap (*because, so, therefore, as, since*). However, interpreting prefers to use *so* over *because*, whereas originals speeches contain more instances of *because* than of *so.*

While we realise that our study has some limitations (we do not consider the impact of the source texts and do not analyse what triggers the explicit use of connectives), we believe that this study will demonstrate the importance of mode, source language, as well as semantics (type of discourse relation) in the analysis of explicitation phenomenon and will be of interest for both translation studies and contrastive linguistics.

References

Blum-Kulka, S. (1986). Shifts of Cohesion and Coherence in Translation. In House, J. and Blum-Kulka, S., editors, *Interlingual and intercultural communication*, pp. 17–35. Gunter Narr, Tubingen.

Crible, L. (2020). Weak and strong discourse markers in speech, chat and writing: Do signals compensate for ambiguity in explicit relations? *Discourse Processes, 57 (9),* pp. 793-807.

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

Gumul, E. (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*, pp. 171–190.

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*, pp. 113–131.

Kajzer-Wietrzny, M. (2012). I*nterpreting universals and interpreting style*. PhD thesis, Uniwersytet im. Adama Mickiewicza, Poznan, Poland. Unpublished PhD thesis.

Karakanta, A., Vela, M., and Teich, E. (2018). Europarl-UdS: Preserving metadata from parliamentary debates. In Fiser, D., Eskevich, M., and de Jong, F., editors, *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Klaudy, K. and Karoly, K. (2005). Implicitation in translation: Empirical evidence for operational asymmetry in translation. *Across Languages and Cultures, 6*, pp. 13–28.

Knaebel, R. (2021). discopy: A Neural System for Shallow Discourse Parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pp. 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Knaebel, R. and Stede, M. (2020). Contextualized Embeddings for Connective isambiguation in Shallow Discourse Parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pp. 65–75, Online. Association for Computational Linguistics.

Lapshinova-Koltunski, E., Pollklasener, C., and Przybyl, H. (2022). Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora. *The Prague Bulletin of Mathematical Linguistics, 119*, pp. 5–22.

Lapshinova-Koltunski, E., Przybyl, H., and Bizzoni, Y. (2021). Tracing variation in discourse connectives in translation and interpreting through neural semantic spaces. In *Proceedings of CODI at EMNLP-2021*, pp. 134–142, Punta Cana and Online. ACL.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). *The Penn Discourse TreeBank 2.0. Technical report*, Linguistic Data Consortium, Philadelphia. Web Download.

Przybyl, H., Lapshinova-Koltunski, E., Menzel, K., Fischer, S., and Teich, E. (2022). EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 1193–1200, Marseille, France. ELDA.

Scholman, M., Dong, T., Yung, F., and Demberg, V. (2021). Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pp. 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Shlesinger, M. (1995). Shifts in cohesion in simultaneous interpreting. *The Translator, 1*, pp. 193–214.

Changing sentiment through professional and student translations

Ekaterina Lapshinova-Koltunski and Pimchanok Sookaim

University of Hildesheim, Germany

KEYWORDS: sentiment, polarity, student translation, professional translation, back-translation

Introduction

We analyse English sentiment (polarity) words in Amazon reviews focusing on their alteration through translation. From the existing studies of sentiment in translation (Lapshinova-Koltunski et al. 2021, Troiano et al. 2020, Mohammad et al. 2016 and Salameh et al. 2015), we know translation can obscure sentiment of the source texts and even change positive and negative aspects. We also know that the level of expertise of translators (professional or novice) may also have an impact on translation of polarity. In this study, we aim to find out if polarity words are added, deleted or altered in translation.

Methodology

We determine the polarity of a piece of text as positive or negative following the approach of a lexicon-based sentiment analysis (Taboada et al. 2011). For this, we use the tool Geist (Kliche, 2020) which integrates SentiWords (Gatti et al., 2016) - a list of weighted negative and positive items. As it is hard to find comparable lists of sentiment words across languages, we decide for analyses in a back-translation setup following (Troiano et al. 2020), so that both the sources and the translations are in the same language.

We select a subset of English Amazon reviews from a publicly available corpus DiHuTra containing translations into three languages by two groups of translators (see Lapshinova et al. 2022). Although reviews are not commonly human-translated, they represent an interesting translation register to analyse, as they are informal which requires creative decisions by translators and would cause variation according to translators' groups. Moreover, they may also reveal some cross-cultural differences that may impact the sentiment. We extract product reviews of the categories *Beauty, Grocery and Gourmet Food, Movies and TV* and *Toys and Games*. Each category contains 14 reviews ranging between 857-1,088 words in length. To obtain the subcorpus of back-translations, we use the open source tool DeepL (accessed on 15.11.2022) to translate the Russian professional and student translations back into English. This results in a dataset with three variants of the same texts: (1) English originals, (2) English machine translations from professional translations and (3) English machine translations from student translations. Note that automatic translation was used to produce comparable texts in English only. Although we are aware of the impact this method of translation may have on the output, we focus rather on the differences between the source translations, i.e. professional vs. student. Therefore, we label our data in the following as originals (orig) – the texts originally written in English, professional translations (prof) – the English texts back-translated from Russian translations by professionals and student translations (stud) – the English texts back-translated from Russian translations by students.

In our analysis, we first compare the overlaps between the texts variants, i.e. how much the back-translated English texts overlap with the English originals. For this, we derive a method used I automatic evaluation of machine translations (BLEU, Papineni et al. 2002). However, we use this method to compare which of the back-translations is more similar to the source text in terms of words and word constructions (n-grams) used in the texts instead of assessing the quality of machine translations. We use the interactive BLEU score evaluator available online (https://www.letsmt.eu/Bleu.aspx). Although the BLEU scores do not deliver any information on the polarity, it give us a generla insight on which translation variant is closer to the original texts. In the next analysis steps, we focus on sentiment. For this, we automatically extract positive and negative words from the originals and both translation variants. After that, we compare the overall sentiment weights and the distribution of positive and negative words

in the subcorpora. We also look into most frequent positive and negative words. In the final step, we qualitatively analyse selected texts to have a better understanding of the observed phenomena.

Analyses and results

The BLEU scores (calculated for each back translation against the original text) show that the difference between the average scores for prof and stud is not big (26.34 and 28.01, respectively). Interestingly, the difference varies across the texts, with the maximum score of 54.29 in prof and 55.57 in stud, and the minimum score of 4.45 and 6.16 respectively. One beauty review achieves 8.75 in prof and 29.9 in stud, which indicates a greater overlap with orig for stud. These results demonstrate that stud translations seem to be closer to the orig texts in terms of word and word construction choice. This may indicate that student translators translate more literally, whereas professionals rather apply a free translation strategy. On the basis of this insights, we assume that stud would be closer to org in terms of sentiment words too.

The results of the automatic sentiment analysis (distribution of polarity words and their weights) are given in Table 1.

Table 1. Quantitative results of sentiment analysis: orig=original English, prof=English back-translated from professional translations, stud=English back-translated from student translations.

subcorpus	total words	positive	positive value	negative	negative value	total value
orig	3735	256	135.23	119	62.53	197.76
prof	3778	237	132.08	129	67.89	199.97
stud	3862	258	138.62	118	64.49	203.11

Table 1 reveals that total values of each subcorpus are similar, with the reviews translated by students having the highest sentimental values. More precisely, stud contains words with stronger sentiment than orig and prof. The ratio between absolute numbers and the values in terms of negative words confirms this: orig contains more negative words, but their polarity is stronger in stud (119 with 62.53 and 118 with 64.49, respectively). Student translations seem to be also more positive than orig and prof, which is similar to the findings in other studies.

Ten most frequently used positive and negative words across the subcorpora overlap. The two top most used positive words (*good* and *great*) of three subcopora are identical. However, there are also positive words in orig that are barely or never used in translations, i.e. *nice* occurs five times in orig, but only once in stud (and never in prof). Translations of *nice* depend on the context of each review and include *great*, *good*, *perfectly*, *beautifully* and *comfortable*. The two top negative words are *fake* and *old*. Interestingly, most slang or colloquial words, such as *knock off*, *back-to-back*, *lame*, reveal variation in their translations: *lame* was translated as *dumb* in prof and as *bullshit* in stud.

In terms of individual categories under analysis, we observe a change from positive to negative in professional translations of beauty reviews: from the positive value of 22.15 and negative value of 18.74 in orig to 21.77 and 21.13 respectively. However, the results of our manual analysis show that this method of automatic sentiment evaluation is not always reliable, as polarity and polarity weights are assigned to the words without considering context, e.g. embedding into a negative construction. Besides that, reviews contain many sarcastic remarks and improvement recommendations with positive words, having, however, an overall negative evaluation. Moreover, some negative and positive words used in our data are not included into the SentiWords list and are not recognised and assigned any values. The manual analysis also reveals that in some cases we do observe changes of sentiment by translators (e.g. *touching* in prof instead of *disturbing* in orig). However, there are also errors that derive from the automatic translation (rare words are wrongly translated).

In our presentation, we will provide more details, illustrate the results and elaborate on the limitations of our approach. We believe that our findings will be of interest for corpus-based translation studies, as they shed light on translation of pragmatic properties.

References

Gatti, L., M Guerini and M. Turchi (2016). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. I*EEE Transactions on Affective Computing, 7 (4)*, 409-421.

Kliche, F. (2020). Die Erschließung heterogener Textquellen für die Digital Humanities. Dissertation, Universität Hildesheim.

Lapshinova-Koltunski, E., M. Popović, and M. Koponen (2022). DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1751–1760, Marseille, France. European Language Resources Association.

Lapshinova-Koltunski, E., F. Kliche, A. Moskvina, and J. Schäfer (2021). Polarity in Translation: Differences between Novice and Experts across Registers. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pp. 66–73, online. Association for Computational Linguistics.

Mohammad, S.M. M. Salameh, and S. Kiritchenko (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research, 55 (1),* 95–130.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Salameh, M., S. Mohammad and S. Kiritchenko (2015). Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the ACL: Human Language Technologies,* pp. 767–777, Denver, Colorado. ACL.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37,* 267–307. Association for Computational Linguistics.

Troiano, E., R. Klinger, and S. Padó (2020). Lost in Back-Translation: Emotion Preservation in Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4340–4354, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fatal attraction – how EU interpreters (more so than translators) are drawn to incorrect renderings when cross-linguistic links are weak. A multifactorial corpus study

Marie-Aude Lefer[1] and Gert De Sutter[2]

[1]UCLouvain, Belgium, [2]Ghent University, Belgium

KEYWORDS: European Parliament, intermodal studies, noun concatenations, Gravitational Pull Hypothesis, regression analysis

In recent years, in the wake of Shlesinger's (1998) pioneering work in corpus-based interpreting studies, corpus translation research has progressively branched out to include intermodal studies, where different mediated language varieties are compared (typically, written translation and simultaneous interpreting; cf. Bernardini et al., 2016). This type of intermodal research has been further promoted by Kotze's (2019) constrained-language framework, which aims to identify the commonalities between language varieties where constraints of different kinds play an above-average role. The key constraint dimension along which translation and interpreting differ is the 'register/modality' dimension (written vs spoken language production).

The present study adds to the growing body of intermodal corpus research by examining the French renditions of English noun concatenations (i.e. sequences of at least two nouns, such as *food prices*) in two modes of interlingual mediation commonly practiced at the European Parliament (EP), namely simultaneous interpreting of the speeches delivered during EP plenary sessions and written translation of the official verbatim reports of these speeches. The reason for examining concatenated nouns is that they have been described as difficult, error-prone items in both modalities. Intermodal research on this topic, however, is still scarce.

In this study, we aim to assess the impact of a large set of frequency-, complexity- and lexicalization-related variables on the use of semantically equivalent (vs non-equivalent) renditions, drawing both on insights from previous empirical research on noun sequences and on Halverson's (2017) cognitive linguistic model of translation, the revised gravitational pull hypothesis. The model posits three cognitive sources of translational effects: source language salience (gravitational pull), target language salience (magnetism) and cross-linguistic link strength (connectivity), where salience is operationalized as, among other things, frequency of use. To date, the model has been tested on a handful of linguistic items, such as morphemes and individual lexemes (Hareide, 2016; Vandevoorde, 2020; Marco, 2021), but it has rarely been used to study items above the word level. However, we believe that the model holds great potential for the study of structures such as concatenated nouns, since psycholinguistic research has shown the crucial role played by frequency in processing and producing these items (Baayen et al., 2010). In addition, to the best of our knowledge the model has not been applied to interpreting, nor has it been used in robust multifactorial research designs such as the one we use here.

The study relies on several corpus frequency counts that function as operationalizations of the three cognitive forces included in Halverson's model, namely gravitational pull (frequency of the concatenation and its constituents in the source language, here English), connectivity (frequency of the cross-linguistic correspondence of the English source concatenation and its French translation or interpretation) and magnetism (frequency of the rendition in the target language, here French). The frequency variables drawn from the gravitational pull model are considered alongside other factors that have been shown to influence compound processing, namely length and lexicalization, with a view to singling out the factors that condition the use of semantically equivalent (vs non-equivalent) renditions in French.

We expect similar trends to emerge in the two modalities, namely that lexicalization, short length and high frequency (reflecting strong gravitational pull, strong connectivity and/or strong magnetism) will go hand in hand with semantically equivalent renditions. However, we expect the effects of these variables to be more visible in interpreting, in view of the fact that "[b]ecause interpreting affords only limited

opportunity for restatement or corrections, it can be seen as the practitioner's default version, with written translation representing a more polished rendition" (Shlesinger & Malkiel, 2005: 185). Contrary to interpreting, written translation is an offline activity, often involving the use of resource tools, self-revision and editorial intervention. In other words, our prediction is that ad hoc concatenations, long concatenations, and concatenations that display low gravitational pull, low connectivity and/or whose equivalents display low magnetism will trigger incomplete renditions, wrong renditions and omissions more frequently in interpreting than in translation.

To test these hypotheses, we make use of corpus data extracted from the *European Parliament Translation and Interpreting Corpus* (EPTIC; Ferraresi & Bernardini, 2019). More specifically, we rely on 106 speeches delivered in English by Members of the European Parliament, commissioners and guests, and their French simultaneous interpretations by highly skilled professionals who are all native speakers of French. We also used the verbatim reports of these speeches and their French translations. The spoken and written components of the subcorpus used in the study each total ca 60,000 tokens. The dataset extracted from this EPTIC subcorpus contains 853 English noun sequences, together with their renditions in interpreted and translated outputs. All occurrences were coded for the response variable 'semantically equivalent rendition' vs 'semantically non-equivalent rendition' and the following explanatory variables: speech-text id, modality (translation or interpreting) and ten gravitational-pull-, magnetism-, connectivity-, complexity- and lexicalization-related variables (in the form of various frequency counts drawn from the Europarl reference corpus and attestedness in several lexicographic and terminological sources). We measured the effect of the explanatory variables on our response variable by means of a generalized linear mixed-effects model (glmm).

A key finding emerging from the regression analysis is that both connectivity and magnetism exert a strong influence on translators' and interpreters' use of semantically equivalent renditions of English concatenated nouns. While highly entrenched cross-linguistic links draw translators and interpreters alike to semantically equivalent renditions, the opposite force is observed in the case of magnetism (target language salience), with strong magnetism leading translators and, more particularly, interpreters to the use of semantically non-equivalent renditions, such as incomplete renditions and wrong renditions. In our presentation, we will discuss these results and their implications for the gravitational pull hypothesis and, more generally, for a usage-based cognitive model of translation.

References

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise, & I. Vogel (Eds.), *Cross-Disciplinary Issues in Compounding* (pp. 257–270). John Benjamins.

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. *Target*, *28*(1), 61–86.

Ferraresi, A., & Bernardini, S. (2019). Building EPTIC: A many-sided, multi-purpose corpus of EU parliament proceedings. In I. Doval, & Nieto, M. T. S. (Eds.), *Parallel Corpora for Contrastive and Translation Studies: New resources and applications* (pp. 123–139). John Benjamins.

Halverson, S. L. (2017). Gravitational pull in translation: Testing a revised model. In G. De Sutter, M.-A. Lefer, & Delaere, I. (Eds.), *Empirical Translation Studies: New methodological and Theoretical Traditions* (pp. 9–45). De Gruyter.

Hareide, L. (2016). Is there Gravitational Pull in translation? A corpus-based test of the Gravitational Pull Hypothesis on the language pairs Norwegian-Spanish and English-Spanish. In M. Ji, M. Oakes, L. Defeng, & Hareide, L. (Eds.), *Corpus Methodologies Explained* (pp. 188–231). Routledge.

Kotze, H. (2019). Converging What and How to Find Out Why. An Outlook on Empirical Translation Studies. In J. Daems, B. Defrancq, & L. Vandevoorde (Eds.), *New Empirical Perspectives on Translation and Interpreting* (pp. 333–371). Routledge.

Marco, J. (2021). Testing the Gravitational Pull Hypothesis on modal verbs expressing obligation and necessity in Catalan through the COVALT corpus. In M. Bisiada (Ed.), *Empirical studies in translation and discourse* (pp. 27–52). Language Science Press.

Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta*, *43*(4). 486–493.

Shlesinger, M., & Malkiel, B. (2005). Comparing modalities: Cognates as a case in point. *Across Languages and Cultures*, *6*(2), 173–193.

Vandevoorde, L. (2020). Semantic differences in translation: Exploring the field of inchoativity. Language Science Press.

Wordless: Integrated corpus tools with multilingual support for the study of language, literature, and translation

Ye Lei

Shanghai International Studies University, China

KEYWORDS: corpus tool, corpus-based studies, corpus linguistics, corpus-based literary studies, corpus-based translation and interpreting studies

This paper presents *Wordless* (Ye, 2023), an integrated corpus tool with multilingual support independently designed and developed by the author for the study of language, literature, and translation, which aims to be a better alternative to other widely-used corpus analysis tools such as *WordSmith*, *AntConc*, *ParaConc*, and *Sketch Engine*.

Although the aforementioned tools have obviously stood to the test of time and facilitated the conduction of many corpus-based studies, recent advancements and bleeding-edge technologies introduced every year in the now rapidly-growing realms of natural language processing (NLP), machine learning, artificial intelligence, and other related areas of study, are not well integrated, and hence under-utilized, during the several-year-long update history of these tools. While computer-savvy researchers have the choice to write their own scripts to process and analyze their corpus data as necessary, the majority of researchers dealing with linguistic data, especially in China where most college students majoring in linguistics, literature, and translation choose to take the liberal-arts exam before entering universities, do not even know where to find the tools and software that they need, let alone make good use of them. And it is not at all painless nor practical for many to learn coding from scratch just in order to try out a new corpus-based approach in their studies. These invisible technology barriers that students and researchers keep facing every day significantly limit the application potentials of corpus-based approaches and techniques in many fields of study, and this is exactly where *Wordless* could come in handy and help bridge the gap.

Wordless is battery-included with a complete set of industrial-strength NLP tools built on bleeding-edge technologies and state-of-the-art research for several dozens of both major and minor languages around the world. Configuration details along with all required data files are automatically handled by the software so that users can dedicate themselves wholehearted to their studies and researches rather than get annoyed by those tricky technical particularities from time to time.

Wordless implements a greater variety of popular statistical methods, including both classical and modern ones, which are the cornerstones that make possible the automation of vocabulary generation and collocation/colligation/keyword extraction. These include 1) dispersion, an indicator of the evenness of word distributions, 2) adjusted frequency, a measurement considering both the raw frequency and distributional evenness of words, which could serve as reliable and easy-to-use references for the work of dictionary/vocabulary/glossary compilers, 3) Bayes factor and effect size, which, when combined with statistics and p-values given by significance tests, could provide stronger and more persuasive evidence for test conclusions, hence helping produce well-founded and more consistent collocation/colligation/keyword lists.

Wordless also comes with a more comprehensive set of built-in data visualization functions, which includes line charts, word clouds, network graphs, and dependency graphs, enabling users to look into their data from different perspectives and brings to them insights and findings that might otherwise be easily overlooked and buried in tabular data.

Other innovative features of *Wordless* include 1) a stand-alone module exclusively used to provide an overview of each corpus file and group all relevant statistics into several tables for further checking by users, 2) a convenience function for the creation of cloze tests, 3) the capability of analyzing both monolingual and parallel corpora, and 4) the ability to parse and visualize syntactic dependency relations, to name just a few. To be specific, the module *Profiler* provides common text statistics, such as type-token

ratios and counts of paragraphs / sentences / sentence segments / words / syllables, and is able to calculate more than a dozen readability measurements which could serve as indicators of text difficulty. The zapping function in the module *Concordancer* can be used to quickly produce a cloze test for testing students' mastery of various inflected forms in the foreign language they are learning. The module *Parallel Concordancer* could be used to produce parallel concordance lines from one-to-one or one-to-many translation corpora or quickly search for additions and deletions in multiple translated texts. The module *Dependency Parser* can be used to analyze dependency relations in every sentence where a given search term is found.

Wordless features a user-friendly and intuitive graphical interface which is specially designed for and exquisitely tailored to the needs of non-technical users, while technical users can benefit from the improved work efficiency at the same time. All functions of *Wordless* can be freely used by anyone for any purpose without registration or subscription. *Wordless* is equipped with full cross-platform support, obviating the need to set up a virtual machine for macOS and Linux users. Lastly, *Wordless* is open-source software and all source codes are hosted on *Github*. Those aspiring to develop future generations of corpus tools are free to view, download, modify, re-distribute the source codes of *Wordless* as long as they abide by the terms of GPLv3 license.

References

Ye, L. (2023). Wordless (Version 3.2.0) [Computer software]. Github. https://github.com/BLKSerene/Wordless

The use of articles in legal translations by Polish advanced learners of English – a corpus based study

Agnieszka Leńko-Szymańska, Łucja Biel, Katarzna Wasilewska

University of Warsaw, Poland

KEYWORDS: articles, legal translation, second language, advanced learners, second language acquisition

The article system is one of the most ubiquitous features of English grammar. Articles are part of a larger class of determiners whose role is to indicate the reference of a noun. They convey the semantic and pragmatic categories of definiteness and specificity, but they also depend on the grammatical category of countability (Ionin, 2006). Yet, in spite of their prominence, the acquisition of articles is particularly hard for EFL learners, especially those whose L1 does not mark definiteness grammatically (Thomas, 1989). The difficulty in acquiring articles has been ascribed to problems in mastering a complex system of grammatical, semantic, and pragmatic relations.

This study examines the use of articles by advanced learners of English whose L1 is Polish, a [-ART] language. The investigation zooms in on one specific register: legal texts. Legal genres provide interesting data to study article use, which so far have not been explored from the acquisitional perspective. Syntactically, legal language is characterized by, among other features, a high degree of nominalization and complex NP structures (Bathia, 1994). At the semantic and pragmatic level, the definiteness and specificity of reference are particularly relevant categories in legal contexts and they need to be expressed precisely and clearly to prevent multiple interpretations. Moreover, the learner legal texts analyzed in this study are translations produced by students in a translation programme. This choice of data elicitation technique combines the advantages of a closed task and of free production, as it reflects a controlled but authentic (even if not typical and common) language use situation. In addition, the application of corpus methodology and tools facilitates quantitative analyses of the data.

The present study addresses the following research questions:

- Is there a difference in the frequency of articles between legal translations produced by Polish advanced learners of English, translations performed by professional translators and comparable non-translated legal texts in English? Is so, does it point to an underuse of articles by learners, observed in earlier studies examining of more common genres (Ekiert, 2004)?
- Is there a variation in the use of articles in L2 learners' texts? If so, does it depend on specific semantic, pragmatic and grammatical contexts? How does it compare with the use of articles in expert translations and in comparable non-translated texts?

The data have been drawn from five corpora: (1) a learner parallel corpus containing 78 Polish-English translations of the same short legal text (PL: statut, EN: articles of association) rendered by MA translation programme students; (2) an expert parallel corpus of translations of the same text produced by 9 experienced Polish legal translators and 1 experienced native-speaking legal translator; (3) a comparable parallel corpus of expert translations containing English translations of Articles of Associations of top 20 Polish listed companies; (4) two comparable corpora of non-translated Articles of Association of top 20 UK and US listed companies as a benchmark.

The numbers of articles have been computed separately for each text in each corpus and than standardized for text length as well as for the number of NPs in the text. Mean frequencies for each corpus are presented in Table 1.

	Learner	Expert Polish	Native	PL20	UK20	US20

	Mean	Sd.	Mean	Sd.		Mean	Sd.	Mean	Sd.	Mean	Sd.
tokens	505.62	19.93	527.56	20.65	496	8922.50	3049.95	33091.40	7824.05	11908.25	7073.47
NPs	131.95	7.00	134.00	7.94	135	2300.95	821.91	7733.50	1850.12	2894.58	1708.21
the_freq	64.92	5.49	63.56	5.81	64	776.05	304.54	2452.65	568.68	912.00	555.80
the_per_100_tokens	12.83	0.89	12.04	0.91	12.9	8.58	1.43	7.42	0.34	7.61	0.64
the_per_100_NPs	49.23	3.95	47.44	3.61	47	33.37	5.45	31.77	1.42	30.93	2.78
a_freq	6.22	2.43	6.33	2.35	6	92.37	32.00	881.40	224.30	227.18	148.06
a_per_100_tokens	1.23	0.48	1.20	0.45	1.21	1.74	0.60	2.69	0.37	1.76	0.49
a_per_100_NPs	4.71	1.88	4.76	1.80	4.44	6.77	2.34	11.53	1.58	7.17	2.09

Table 1: (Relative) frequencies of the definite and indefinite articles in the five corpora

Results of ANOVA tests and pos-hoc analyses reveal statistically significant differences between the five corpora. Yet, the observed differences are contrary to the hypotheses made an the outset of the study. The learner translations display exceptionally high frequencies of the definite article (Mthe/tokens=12.83 and Mthe/NPs=49.23) and low frequencies of the indefinite article (Ma/tokens=1.23 and Ma/NPs=4.76) in comparison with the benchmark established by the English non-translated texts. However, the learner frequencies are not different than the respective frequencies in the expert translations of the same text, including the rendition produced by the native-speaking legal translator. These surprising results will be discussed in the context of an interplay between the characteristics of the text and more general features of so-called translationese (Baker et al., 1993).

Next, the most frequent NP heads in the learner corpus were examined for the occurrence of articles. The analysis demonstrate different degrees of variation in article use, which can be linked to the categories of definiteness, specificity and countability/grammatical number displayed by the referents denoted by the head nouns. In the case of nouns indicating a single and unique referent (company, board) the learners are almost 100% unanimous in the choice of the definite article. More variation can be observed in the case of referents which are definite but not specific and are represented one or more entities (member, shareholder, resolution). Another type of referent which was analyzed involves a noun denoting an activity (vote). Finally, nouns indicating quantity (majority, number) were considered. In all cases, learner solutions were compared to expert translations and similar noun phrases in non-translated texts.

Overall, both the analysis of the frequencies of articles in the five corpora as well as the examination of the variation of article use in most frequent noun phrases paint an interesting picture the acquisition of the article system by Polish advanced learners of English, which will be discussed in the presentation

References

Baker, M., Francis, G. & Tognini-Bonelli, E. (1993). *Corpus Linguistics and Translation Studies: Implications and Applications*. Amsterdam: John Benjamins.

Bhatia, V. K. (1994). Analysing Genre: Language Use in Professional Settings. New York: Longman.

Ekiert, M. (2004). Acquisition of the English article system by speakers of Polish in ESL and EFL settings. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *4*(1): 1–23.

Ionin, T. (2006). This is definitely specific: specificity and definiteness in article systems. *Natural Language Semantics*, *14*(2): 175–234. http://www.jstor.org/stable/23749620

Thomas, M. (1989), The acquisition of English articles by first- and second language learners. *Applied Psycholinguistics, 10*: 335–355.

Adjective position and Machine-Translationese: An insight from the COVALT corpus

Josep Marco and Maria Ferragud

Jaume I University, Spain

KEYWORDS: adjective position, Machine-Translationese, English-Catalan, COVALT, Gravitational Pull Hypothesis

Toral (2019) argued that features normally associated with translated language are exacerbated in post-edited translations. He coined the term *post-editese* (in an analogy with *translationese*) to refer to this phenomenon. If Toral's claim is true, it can be further claimed that such an exacerbation must still be more noticeable in raw Machine Translation (MT). Toral drew on evidence from so-called document-level metrics, including lexical variety, lexical density, length ratio and part-of-speech sequence. Several studies within the COVALT group still unpublished so far – e.g. Marco (2021), Marco and Oster (2022) – have attempted to put to the test what might be called the Machine-Translationese (MTese) hypothesis – basically, Toral's post-editese hypothesis applied to the raw output of MT – on the basis not of document-level metrics but of particular linguistic elements or constructions. Marco (2021) found evidence of MTese in the Catalan modal indicator *caldre* and in the imperfective/perfective aspectual distinction. In Marco and Oster (2022), though, constructions conveying passive construal were used as indicators and yielded more blurred results, as exacerbation was observed for some constructions and language pairs, but not for others.

The present study aims to test out the MTese hypothesis on adjective position in Catalan. Catalan, like other Romance languages, may place attributive adjectives either before or after the noun, the difference between both constructions being mostly stylistic in nature. Noun + Adjective (NA) tends to be the unmarked order, whereas the opposite is mostly used in non-specifying contexts, i.e. when the quality or relation conveyed by the adjective is not liable to make distinctions among members of a category – e.g. "grans praderies" ('great prairies') does not distinguish between great and small, but assumes that *all* prairies are great. On the other hand, English, like other Germanic languages, places attributive adjectives in pre-nominal position as a rule, with very few exceptions.

The present study is inspired by Oster (2019, 2020), who conducted an ambitious analysis of adjective position in a number of language combinations. Her 2019 study was developed within the framework of Halverson's (2003, 2017) Gravitational Pull Hypotheses, a cognitively-oriented theory which envisages three causes of translational effects: target language influence (or *magnetism*), source language influence (or *gravitational pull* proper) and nature and strength of the links between source and target language nodes (or *connectivity*). Over- or under-representation of a given feature in translations (as compared to non-translations in the target language) will be conditioned by the specific configuration of these three causes in a specific language pair. In Oster's 2020 study, though, the GPH connection was dropped.

Oster's focus was on the translation/non-translation dimension. In the present study we intend to add a third column to her kind of analysis and compare both non-translations (CA) with (human) translations (HT) and human with machine translations (MT). We will draw on the English-Catalan sub-corpus of COVALT plus the added component of machine translations of the same English source texts. This sub-corpus corpus can be visited at <http://150.128.97.141/cqpweb/>. The original COVALT corpus was made up of Catalan translations published in the region of Valencia between 1990 and 2000 of narrative works written in English, French and German. That original corpus was expanded in different directions. The three corpus components under scrutiny have the following size: 1,551,521 words (CA), 1,335,134 words (HT) and 1,291,458 words (MT). All texts are lemmatised and pos-tagged with FreeLing. Corpus analysis is carried out with CQPweb.

Our methodology proceeds in three steps. The three components just mentioned (CA, HT and MT) are first queried for the two constructions NA and AN. Secondly, since the total number of matches for these

queries is unmanageable, they are thinned to 5% and manually sifted in order to discard false positives. Finally, results are normalised per 100,000 words (as the three corpus components differ in size) for ease of comparison, and differences are tested for statistical significance through chi-square. Table 1 shows the results of this analysis. The second column offers the number of true positives, and the third indicates relative frequencies. The fourth column is a projection of the results in the second column over the whole number of occurrences, assuming that those results are representative. The fifth column displays normalised frequencies per 100,000 words.

Table 1 Result of the analysis of the constructions NA and AN in CA, HT and MT (English-Catalan)

	Raw frequency (in 5% random sample)		Relative frequency (in 5% random sample)		Projection		Normalised frequency (100,000 words)	
	NA	AN	NA	AN	NA	AN	NA	AN
CA	1,027	488	67.79%	32.21%	20,540	9,760	1,323.86	629.06
HT	920	525	62.90%	37.10%	18,400	10,500	1,378.14	813.03
MT	997	390	71.88%	28.12%	19,940	7,800	1,543.99	603.97

These results show that the NA construction is under-represented in HT when compared to non-translations (χ^2=5.579, df=1, p<0.025, the significance threshold standing at 3.84 for p<0.05). This means that, in our English-Catalan human translations, gravitational pull prevails over magnetism. However, there is no evidence for exacerbation of this particular feature in MT. In fact, the distribution of the two constructions NA and AN is closer to CA than it is to HT. The difference in distribution between the two constructions in HT and MT is statistically significant (χ^2=21.832, df=1, p<0.001), but points in the opposite direction to the difference between CA and HT. Therefore, for the particular feature adjective position, what MTs in our corpus exacerbate is not the translational effects subsumed under the heading *translationese*, but features defining Catalan non-translations (again, the difference in distribution between the two constructions in CA and MT is statistically significant, with χ^2=5.748 , df=1 and p<0.025). In other words, Catalan non-translations (CA) stand somewhere in between HTs, where gravitational pull prevails over magnetism, and MTs, where magnetism prevails over gravitational pull.

Our study will include a qualitative study of query matches for both NA and AN in order to determine what factors favour the use of either construction. Oster (2019) mentions the following factors, among others: type of adjective, post-modification of the adjective, post-modification of the noun, multiple noun modification in the source text and degree of fixation of a particular NA/AN construction.

References

Halverson, S. (2003). The cognitive basis of translation universals. *Target*, 15(2), 197–241.

Halverson, S. (2017). Developing a cognitive semantic model: magnetism, gravitational pull, and questions of data and method. In G. De Sutter, M. A. Lefer & I. Delaere (Eds.), *Empirical Translation Studies. New Methods and Theoretical Traditions* (pp. 9–45). Mouton de Gruyter.

Marco, J. (2021, November 4–5). *Machine-translationese in the English-Catalan sub-corpus of COVALT* [Conference presentation]. PETRA-E Conference 2021: Literary Translation Studies Today & Tomorrow, Trinity College Dublin, Ireland.

Marco, J. & Oster, U. (2022, September 22-23). *Passive construal as testing ground for the Gravitational Pull Hypothesis and Machine-translationese in the COVALT corpus* [Conference presentation]. Translation in Transition 6, Charles University, Prague, Czech Republic.

Oster, U. (2019). Anteposició i posposició de l'adjectiu en textos originals i traduïts en català i en castellà. In T. Molés-Cases & M. D. Oltra Ripoll (Eds.), *El corpus COVALT: model de llengua, sociologia del traductor i anàlisi traductològica* (pp. 117–139). Shaker Verlag.

Oster, U. (2020). Sobrerrepresentación de la anteposición del adjetivo en textos traducidos - ¿realidad o prejuicio? In M. A. Recio Ariza, S. Roiss, B. Santana, I. Holl, M. De la Cruz Recio &

P. Zimmermann (Eds.), *Del texto a la traducción. Estudios en homenaje a Pilar Elena* (pp. 115–132). Comares.

Toral, A. (2019). *Post-editese: an exacerbated translationese*. arXiv:1907.00900.Čermáková, A. (2015). Repetition in John Irving's novel *A Widow for One Year*. A Corpus stylistic approach to literary translation. *International Journal of Corpus Linguistics*, 20(3), 355–77.

.

Repetition in English novels and their Italian and Polish translations: A Multifactorial Study

Lorenzo Mastropierro[1], Łukasz Grabowski[2]

[1]University of Insubria, Italy, [2]University of Opole, Poland

KEYWORDS: repetition, reporting verbs, literary translation, multiple regression

The exploration of repeated patterns in language is the foundation of much corpus research, including corpus-based translation and interpreting studies. The underlying assumption is that the repetition of a linguistic feature is a reflection of its functional relevance (Mahlberg 2010: 297). Studying repeated patterns can therefore shed light on how texts enact their communicative functions and convey meaning. In the context of translation studies, examining repetition has a twofold value: it encompasses the exploration of both the role of repetition in the source text (ST) and its reproduction in the target text (TT). The relationship between repetition in the ST and its reproduction in the TT is far from straightforward, as it can be affected by factors such as differences in stylistic conventions between languages and cultures (Piotrowski 1994; Nádvorníková 2020), as well as typological divergences between source and target language (Comrie 1981). Despite its importance and complexity, there is surprisingly little research that explores repetition in corpus-based translation studies. Most of the existing literature approaches repeated patterns as a way to study aspects of language other than repetition itself. Only a minority of studies aim instead to investigate directly how the repetition of functionally relevant features is dealt with in translation. Insights from translation theory suggest that "[o]ne of the most persistent and inflexible norms in translation […] is that of avoiding repetition" (Ben-Ari 1998: 3), and the existing corpus-based evidence has so far aligned with that. Studies such as Mastropierro (2020) and Čermáková (2015, 2018) show that translators tend to avoid lexical repetition in favour of variation. These studies, though, tend to focus on the consequences of repetition avoidance on the style and interpretation of individual TTs, rather than on repetition as a linguistic and translational phenomenon in its own right. Hence, more research that explores repetition in translation – as opposed to its consequences – on a wider variety of texts and language pairs is needed. This paper contributes to redressing the gap through a study of lexical repetition, more specifically the repetition of English reporting verbs in 11 novels and their translations in Italian and Russian. Capitalising on recent methodological developments in corpus-based translation research (e.g. Kruger & van Rooy 2016; DeSutter and Lefer 2020; Kajzer-Wietrzny & Grabowski 2021; Calzada Pérez & Laviosa 2021; Mastropierro 2022), we use statistical multivariate methods (multiple regression) in order to explore the effect that multiple factors may have on the reproduction or avoidance of repetition in translation, providing in this way important insights into how the translation of functionally relevant repeated items is dealt with.

We focus on the repetition of reporting verbs because they enact an important stylistic function in literary texts; that is, reporting verbs play a major role in the characterisation process (Culpeper 2001; Ruano San Segundo 2016, 2017; Eberhardt 2017; Mastropierro forthcoming). The repetition of reporting verbs has functional and stylistic relevance (Nádvorníková 2020), and alterations to reporting verbs patterns in translation has been shown to affect character development (Mastropierro 2020; Čermáková & Mahlberg 2018). Reporting verbs are retrieved from *InterCorp v. 15* (Čermák & Rosen 2012), a multilingual parallel corpus which includes – among many other text types and language pairs – English novels and their translations in Italian and Polish. We identified all the English novels with an available translation in Italian and Polish, and then selected texts for which we could retrieve at least 500 reporting verbs per language, resulting in 11 English novels by five dierent authors and their translations into our two target languages. We fit multiple regression models to assess the effect that four predictor variables have on our output variable, that is, the number of different target language verb types a ST reporting verb is translated into. In other words, we aim to explain what factors can impact on the reproduction or the avoidance of reporting verbs repetition in our TTs. The first variable is the frequency of the ST verb: how many times it occurs in each ST. The second factor is the polysemy of the ST verb: how many different meanings the ST verb can have, retrieved from *WordNet 3.1* (Fellbaum 1998). The third factor is the number of the ST verb

translation equivalents: how many different ways the ST verb can be translated into the target language, retrieved from *Treq* (Vavřín and Rosen 2015). The fourth factor is the type of ST reporting verb, based on Caldas-Coulthard's (1987) taxonomy, which groups reporting verbs in seven semantic and/or functional categories.

This analysis will allow us to answer the following research questions:

- RQ1: What linguistic factors have a significant effect on the avoidance or reproduction of reporting verbs repetition in the Italian and Polish novels?

- RQ2: Are there text- or language-specific differences in the way these factors impact reporting verbs repetition in the TTs and/or language pairs taken into account? Or can a more generalisable trend in how translators deal with repetition be recognised?

By answering these questions, this study will offer important insights into the translation of repetition. A data-based and multifactorial picture of when repetition is maintained or avoided in translation can have significant implications in the context of professional practice, which can feed into the discussion, and improvement, of translation training. By learning what can prompt the translation of repeated reporting verbs into a wider lexical variety, we can improve translation strategies to deal with repetition, making them more sensitive to the stylistic effects that repetition can enact.

References

Ben-Ari, N. (1998). The ambivalent case of repetitions in literary translation. Avoiding repetitions: A 'universal' of translation. *Meta*, 43(1), 68–78.

Caldas-Coulthard, C. R. (1987). Reported speech in Written narrative texts. In *Discussing Discourse*, edited by M. Coulthard, 149–67. Birmingham: University of Birmingham.

Calzada Pérez, M., & Laviosa, S. (2021). Twenty-five years on: Time to pause for a new agenda for CTIS. *MonTI*, 13, 7–32.

Comrie, B. (1981/1989). Language Universals and Linguistic Typology: Syntax and Morphology. Oxford: Basil Blackwell.

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), 411–27.

Čermáková, A. (2015). Repetition in John Irving's novel *A Widow for One Year*. A Corpus stylistic approach to literary translation. *International Journal of Corpus Linguistics*, 20(3), 355–77.

Čermáková, A. (2018). Translating children's literature: Some insights from corpus stylistics. *Ilha Desterro*, 71(1), 117–33.

Čermáková, A., & Mahlberg, M. (2018). Translating fictional characters – Alice and the Queen from the Wonderland in English and Czech. In *The Corpus Linguistics Discourse. In Honour of Wolfgang Teubert*, edited by A. Čermáková and M. Mahlberg, 223–53. Amsterdam and Philadelphia, PA: John Benjamins.

Culpeper, J. (2001). Language and Characterisation: People in Plays and Other Texts. Harlow: Pearson Education.

De Sutter, G., & Lefer, M.-A. (2019). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, 28(1), 1–23.

Eberhardt, M. (2017). Gendered representations through speech: The case of the Harry Potter series. *Language and Literature*, 26(3), 227–46.

Fellbaum, C., (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Kajzer-Wietrzny, M., & Grabowski, Ł. (2021). Formulaicity in constrained communication: An intermodal approach. *MonTI*, 13, 148–83.

Kruger, H., & Van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, 37(1), 26–57.

Mahlberg, M. (2010). Corpus linguistics and the study of nineteenth-century fiction. *Journal of Victorian Culture*, 15(2), 292–98.

Mastropierro, L. (2020). The translation of reporting verbs in Italian: The case of the Harry Potter series. *International Journal of Corpus Linguistics*, 25(3): 241–69.

Mastropierro, L. (2022). The avoidance of repetition in translation: A multifactorial study of repeated reporting verbs in the Italian translation of the Harry Potter series. In *Advances in Corpus Applications in Literary and Translation Studies*, edited by L. Defang & R. Moratto, 138-157. London/New York: Routledge.

Mastropierro, L. (forthcoming). Gendered voices in translation: Reporting verbs in the Italian translation of the Harry Potter series. In *Good Girls and Brave Boys: 19th Century and Contemporary Children's Literature and Childhood*, edited by M. Mahlberg & A. Čermáková. London: Bloomsbury.

Nádvorníková, O. (2020). Differences in the lexical variation of reporting verbs in French, English and Czech fiction and their impact on translation. *Languages in Contrast*, 20(2), 209–34.

Piotrowski, T. (1994). *Problems in bilingual lexicography*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Ruano San Segundo, P. (2016). A corpus-stylistic approach to Dickens' use of speech verbs: Beyond mere reporting. *Language and Literature*, 25(2), 113–29.

Ruano San Segundo, P. (2017). Reporting verbs as a stylistic device in the creation of fictional personalities in literary texts. *Journal of the Spanish Association of Anglo-American Studies*, 39(2), 105–24.

Vavrín, M., & Rosen, A. (2015). "Treq (v. 2.1)." https://treq.korpus.cz/. Accessed Jun 2023.

Sketch Engine as a tool for translators and interpreters

Ondřej Matuška

Sketch Engine

KEYWORDS: corpus tools, corpora

This demo aims to showcase the capabilities of Sketch Engine in the field of terminology extraction and usage checking, enabling users to enhance the linguistic quality of their translations. With support for over a hundred languages, Sketch Engine offers a versatile platform for linguistic analysis. Users have the option to explore the preloaded linguistic data, known as corpora, or upload their own documents for in-depth examination.

Sketch Engine stands as an online text analysis tool specifically designed to handle large sets of language samples, known as text corpora. Its primary objective is to discern the typical and frequent linguistic elements within a given language, while also identifying rare, outdated, or emerging words, grammar, and phrases. As a result, Sketch Engine provides an exceptional means to comprehend the inner workings of language. Its advanced algorithms meticulously analyze billions of words across diverse text corpora to instantaneously determine the norms, outliers, and emerging trends in language usage. Furthermore, Sketch Engine serves as an invaluable resource for applications such as text analysis and text mining.

Linguists, lexicographers, translators, students, and teachers all benefit from the capabilities of Sketch Engine. Moreover, it has become the preferred choice for publishers, universities, translation agencies, and national language institutes worldwide. Sketch Engine boasts an impressive collection of 600 readily available corpora in over 100 languages. Each corpus encompasses up to an astounding 60 billion words, ensuring that users have access to a truly representative sample of language. This vast and diverse linguistic dataset enables thorough analysis and facilitates the exploration of language nuances across various domains.

The user-friendly interface of Sketch Engine empowers individuals and organizations to harness the power of language analysis with ease. Whether it's identifying domain-specific terminology, scrutinizing the usage of words and phrases, or tracking language evolution, Sketch Engine provides indispensable insights and tools. By leveraging its extensive corpora and advanced algorithms, users can significantly enhance the accuracy and fluency of their translations, thereby elevating the overall linguistic quality.

In conclusion, Sketch Engine serves as a comprehensive online tool for terminology extraction, usage checking, and linguistic analysis. Its wide range of features, extensive linguistic corpora, and user-friendly interface make it an invaluable asset for linguists, translators, educators, and language enthusiasts worldwide. With Sketch Engine, users can delve into the intricacies of language, identify emerging trends, and ultimately achieve greater proficiency in their linguistic endeavors.

Syntactic complexity in Czech, French and English fiction and non-fiction: A pilot study using the Universal Dependencies annotation

Olga Nádvorníková

Charles University, Czechia

KEYWORDS: syntactic complexity, Universal Dependencies annotation, cross-linguistic differences, fiction, non-fiction, multilingual corpora

The present paper aims at investigating the possibilities of research into cross-linguistic differences in syntactic complexity using syntactic annotation according to the project of Universal Dependencies (UD, see de Marneffe et al. 2021). On a pilot study of English, French and Czech translated and non-translated fiction and non-fiction texts, we examine the potential application of this approach in contrastive and translation studies and the possibilities of its implementation into the interface of the InterCorp multilingual corpus (https://intercorp.korpus.cz/). For the purpose of the study, we define the syntactic complexity in terms of *the number and variety of elements and the elaborateness of their interrelational structure* (Rescher 1998: 1). We argue that given the structural differences and the differences in stylistic traditions, languages vary as to the preferred means of information packaging in sentences (see e.g. Cosme 2006 and 2008, Fabricius-Hansen 1998 and 1999, Solfjeld 1996, Chuquet & Paillard 1987), i.e. also in the degree of their syntactic complexity (Szmrecsanyi 2004, etc.). These differences may be quantified using specific syntactic complexity measures, which allows for cross-linguistic comparison.[1]

Previous research, focussed especially on language proficiency assessment, has elaborated numerous measures of syntactic complexity of sentences (cf. 130 different measures in the survey in Jagaiah et al. 2020). As shown in Nádvorníková (2020), however, some of these measures are not reliable cross-linguistically: for example, the traditional sentence length measure calculated in number of words cannot be used while comparing typologically different languages, such as analytic French and synthetic Czech. Thus, in our research, we opted for two measures based not on the number of words, but on the number and hierarchical relations of clauses: subordination ratio and (maximum) syntactic depth (*average maximum depth of embedded subordinate clauses*). Subordination ratio is calculated as the number of T-units (see Hunt 1965) + the number of dependent clauses / the number of T-units (Jagaiah et al. 2020: 2600); the maximum syntactic depth is given by the number of clauses that are sequentially nested inside one another.

These measures are calculated on data extracted from the InterCorp parallel corpus (Čermák & Rosen 2012 and Rosen, Vavřín & Zasina 2021), recently annotated in UD (de Marneffe et al. 2021).[2] InterCorp is a large multilingual corpus including (in its UD-annotated release 13ud, https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze13ud) 37 languages and 1,7 billion of words in

[1] Quantification of various syntactic complexity measures (SCMs) based on UD annotation is available in *Linguistic Profiling* online tool as well (http://linguistic-profiling.italianlp.it/, see Dell'Orletta et al. 2011). For our research, however, we found the tool's results to be unreliable since it fails to consider the conj deprel (conjuncts). In consequence, it considerably overestimates the number of subordinate clauses and underestimates the number of roots (T-units). Consequently, we decided not to use it in our research, but the tool can serve as inspiration for SCMs available via UD. (For a similar tool in Polish see Gruszczyński & Ogrodniczuk 2015).

[2] The calculation of the subordination ratio was carried out directly in the corpus interface, using the CQL query `[deprel="root"|(deprel="conj" & p_deprel="root")]` for T-units and `[deprel="csubj|ccomp|advcl|acl|acl:relcl|parataxis|xcomp"|(deprel="conj" & p_deprel="csubj|ccomp|advcl|acl|acl:relcl|parataxis|xcomp")]` for subordinate clauses. The Max.Tree.Depth, however, was computed using a script in Perl (I would like to express my gratitude to Martin Vavřín from the Institute of the Czech National Corpus for developing the script).

various text types (fiction, non-fiction, journalistic texts, Acquis Communautaire, *Bible* and EuroParl).[3] For the current study, we have focused on the fiction and, to some extent, to non-fiction, because of the high quality of translations and alignment in these sub-corpora. The pilot study on the syntactic complexity is conducted on two levels of granularity: (1) analyzing the whole sub-corpora across the three languages (translated and non-translated, fiction and non-fiction),[4] and (2) performing a detailed analysis of six specific texts available in the three languages, namely *L'Etranger* and *La Peste* by Albert Camus, and *Žert* and *Směšné lásky* by Milan Kundera, along with two non-fiction texts from the domain of economy).

Table 1. Size of the fiction sub-corpora of the InterCorp multilingual corpus (in tokens). EN, CS, FR = source languages, en, cs, fr = target languages (translations between French and English, marked in italics, were not included in the analysis, as the subcorpora are too small).

Czech	Nº of tokens	English	Nº of tokens	French	Nº of tokens
CS(orig)	19,429,751	EN(orig)	24,329,985	FR(orig)	8,597,632
cs(FR)	8,481,360	en(CS)	4,669,877	fr(CS)	5,112,428
cs(EN)	32,329,585	*en(FR)*	*1,256,457*	*fr(EN)*	*2,363,498*

At the first level of analysis, we explore potential applications of this approach in contrastive linguistics, translation studies, and register variation analysis. Using the contrastive approach, the most prominent in this study, we compare the source text subcorpora in Czech, English and French fiction (the 1st line in Table 1), using the two syntactic measures. On the basis of previous research (Čermák et al. 2020, Malá & Šaldová 2015, Nádvorníková 2017 and 2021, etc.), we assume that French texts will exhibit the highest degree of syntactic complexity, while the Czech texts the lowest, with English texts falling in between. The preliminary results confirm this hypothesis, as the subordination ratio is 2.10 for French, 1.96 for English and only 1.63 for Czech. The differences are statistically significant at $p < 0.0001$, according to a chi-squared test. However, the effect size, measured as Cohen's h, is small: the lowest for EN-FR (0.06), the highest for comparison with Czech: CS-EN (0.21), and CS-FR (0.25). These results may reflect the high stylistic variability in fiction, where the syntax can range within the same language from intentionally very simple (e.g. in *L'Etranger* by Albert Camus) to highly hierarchical (e.g. in *A la recherche du temps perdu* by Marcel Proust). For further research, it may be more appropriate to analyze carefully constructed comparable subcorpora (in contrastive research) or focus on individual texts (in translation studies).

Regarding the potential application in translation studies, we contend that even though the aforementioned contrastive differences are small, they can still have an impact on the syntactic complexity of target texts, due to cross-linguistic interference, and cause differences between translated and non-translated texts in the same language. For fr(CS) and for cs(FR), the results support the hypothesis, as the subordination ratio is lower in fr(CS) than in FR and higher in cs(FR) than in CS. However, in en(CS), the result contradicts the hypothesis, as the subordination ratio is higher in en(CS) than in EN. Moreover, in all the three combinations, the Cohen's values are extremely low: 0.05 for cs(FR), 0.04 for en(CS) and 0.03 for fr(CS)).

As for the register variation analysis, we compared the syntactic complexity of non-translated fiction in the three languages with non-translated non-fiction (2,383,843 tokens for French, 2,224,265 for English and 887,089 for Czech). On the basis of previous studies (Cvrček et al. 2020, etc.), we hypothesize that

[3] The reliability of the syntactic annotation (*label attachment score*) is quite high in all the three linguistic subcorpora under analysis: cs: 94.23% (cs_fictree), en: 91.19% (en_gum), fr: 92.71% (fr_gsd). https://github.com/ufal/udpipe/blob/udpipe-2/models-2.10/EVALUATION-logs.txt

[4] Since Czech is the pivot language of the corpus (see Nádvorníková 2016), all texts are available in this language; the intersections between other languages are more limited, however, for the three languages under analysis in this study, they represent more than 1 million of tokens (FR-cs-en 1,297,024, EN-cs-fr 2,191,009 and CS-fr-en 1,710,145). In this study, the intersections of the three languages have been used only in the analysis of six specific texts.

the degree of syntactic complexity will be higher in non-fiction than in fiction. The results confirm this hypothesis, the respective subordination ratios being 1.74 for Czech, 2.49 for English and 2.83 for French, the Cohen's h 0.10 for EN-FR, 0.31 for EN-CS and 0.41 for CS-FR. At the same time, these results corroborate the afore-mentioned differences between the three languages.

The aim of the second level of the analysis is twofold: (1) verify the reliability of the analysis through manual checks of the parallel occurrences, (2) explore the potential application of the approach in contrastive analysis, focusing on individual parallel sentences that exhibit differences in syntactic complexity and analyzing the potential triggers of the changes. The analysis has revealed that certain observed differences in syntactic complexity between individual sentences are due to technical reasons (e.g. in some texts, the semi-colon is considered a final mark of the sentence) or to language-particular differences in annotation of the same linguistic phenomena (e.g. the clausal `acl` deprel in French often corresponds to non-clausal `amod` deprel in Czech). However, most of the differences in syntactic complexity reflect structural differences between the analyzed languages, shifts resulting from specific features of translation, particularly normalization, or divergences in stylistic traditions.

References

Čermák, Petr et al. (2020). *Complex Words, Causatives, Verbal Periphrases and the Gerund: Romance Languages Versus Czech (A Parallel Corpus-Based Study)*. Praha: Karolinum. Available at: http://hdl.handle.net/20.500.11956/117388

Čermák, František & Rosen, Alexandr (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), 411–427.

Chuquet, Hélène & Paillard, Michel (eds). (1987). *Approche linguistique des problèmes de traduction anglais >< français*. Paris: Gap / Ophrys.

Cosme, Christelle (2006). Clause combining across languages. A corpus-based study of English-French translation shifts. *Languages in Contrast*, 6(1), 71–108.

Cosme, Christelle (2008). A corpus-based perspective on clause linking patterns in English, French and Dutch. In Cathrine Fabricius-Hansen & Wiebke Ramm (eds.). *"Subordination" versus "Coordination" in sentence and text: a cross-linguistic perspective*, 89–114. Amsterdam: John Benjamins.

Cvrček Václav, Zuzana Laubeová, David Lukeš, Petra Poukarová, Anna Řehořková, Adrian Jan Zasina (2020). *Registry v češtině* [Registers in Czech]. Praha: Nakladatelství Lidové noviny.

Dell'Orletta F., Montemagni S., Venturi G. (2011). "*READ-IT: assessing readability of Italian texts with a view to text simplification".* In: SLPAT '11 – SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (Edimburgo, UK, 30 Luglio 2011). Proceedings, 73–83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

Fabricius-Hansen, Cathrine (1998). Informational density and translation, with special reference to German-Norwegian-English. *Language and Computers*, 24, 197–234.

Fabricius-Hansen, Cathrine (1999). Information packaging and translation: aspects of translational sentence splitting (German–English/Norwegian). In Monika Doherty (ed.), *Sprachspezifische Aspekte der Informationsverteilung*, 175–214. Berlin: Akademie Verlag.

Gruszczyński, Włodzimierz & Ogrodniczuk, Maciej (2015). *Jasnopis czyli mierzenie zrozumiałości polskich tekstów użytkowych*. Warszawa: SWPS Uniwersytt humanistycznospołeczny.

Hunt, K.W. (1965). Grammatical structures written at three grade levels (Research Report No. 3). Available at: https://eric.ed.gov/?id=ED113735.

Jagaiah, Thilagha, Olinghouse, Natalie G. & Kearns, Devin M. (2020). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. *Read Writ* 33, 2577–2638. https://doi.org/10.1007/s11145-020-10057-x

de Marneffe, Marie-Catherine, Manning, Christopher D., Nivre, Joakim, and Zeman, Daniel (2021). Universal Dependencies. *Computational Linguistics*, 47(2), 255–308.

Nádvorníková, Olga (2016). Le corpus multilingue InterCorp et les possibilités de son exploitation, in: Éva Buchi, Jean-Paul Chauveau & Jean-Marie Pierrel (ed.) *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15–20 juillet 2013)*. Strasbourg: Société de linguistique romane/ÉLiPhi. http://www.atilf.fr/cilpr2013/actes/section-16.html

Nádvorníková, Olga (2017). Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus. In: *Language Use and Linguistic Structure*. Olomouc: UPOL, 445–461. http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016-proceedings.pdf

Nádvorníková, Olga (2020). The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: a question of methodology. *Linguistica Pragensia*. 30(2), 30-50. ISSN 1805-9635. Available at: doi:10.14712/18059635.2020.1.2

Nádvorníková, Olga (2021). Contexts and Consequences of Sentence Splitting in Translation (English-French-Czech). *Research in Language*, 19(3), 229-250. Available at: https://czasopisma.uni.lodz.pl/research/issue/view/1045

Rescher, Nicholas (1998). *Complexity: A Philosophical Overview*. New Brunswick NJ: Transaction.

Rosen, Alexandr, Vavřín, Martin, and Zasina, Adrian Jan (2021). InterCorp corpus – Czech, version 13ud from 22/12/2021. Prague: Institute of the Czech National Corpus. Available at: https://kontext.korpus.cz/

Solfjeld, Kåre (1996). Sententiality and translation strategies German-Norwegian. *Linguistics*, 34, 567–590.

Szmrecsanyi, Benedikt (2004). On operationalizing syntactic complexity. In *Le poids des mots*. Proceedings of the 7th International Conference on Textual Data Statistical Analysis Louvain-la-Neuve, March 10–12, 2004, Vol. 2, Gérard Purnelle, Cédrick Fairon & Anne Dister (eds), 1032–1039. Louvain-la-Neuve: Presses Universitaires de Louvain.

Cognitive and linguistic factors influencing the French translation of English noun sequences: Evidence from a multiple translation corpus

Thomas Prinzie

UCLouvain, Belgium

KEYWORDS: Empirical Translation Studies, Translation variability, Gravitational pull model, Salience, Noun sequence

It has been suggested that the field of Empirical Translation Studies lags behind in theory development and would benefit from improved interaction with cognitive linguistics (De Sutter & Lefer, 2020; Halverson & Kotze, 2021). While most translation research to date on cognitive processes has been conducted using process methods, there has been growing interest in applying corpus-based methods as well. In this context, Halverson's (2017) revised "gravitational pull" model has set out to explain translators' behavior in terms of usage-based factors such as the salience of source and target items and the entrenchment of translation pairs. Halverson's model has been further tested and extended in several recent studies (e.g., Lefer & De Sutter, 2022; Heilmann et al., 2022). Still, there remain unresolved issues regarding how these factors are best operationalized, due in part to the rather confounded treatment of the construct of salience in the literature (Boswijk & Coler, 2020) and the realization that frequency data alone may not fully capture its effects (Divjak, 2019).

It may also be helpful to distinguish the two modes of processing involved: (1) a fast, largely automatized behavior involving direct one-to-one links from a source language (SL) form to a target language (TL) form; and (2) a slower behavior involving conscious decisions regarding potentially multiple interpretations (from SL form to meaning) and realizations (from meaning to a TL form). De Groot (1997) termed these two modes *horizontal* and *vertical* translation. One view holds that the former is used by default whenever possible and that switching to the latter (also called 'monitoring') is triggered by a perceived difficulty; Halverson (2019) thus favors the terms *default translation* and *conscious deliberation*. Surprisingly, there have been few empirical investigations until recently. I argue that considering them separately may help to disentangle distinct factors contributing to translator choices: (1) in horizontal mode, the presence of entrenched solutions versus potential SL interference; and (2) in vertical mode, the influence of semasiological salience on interpretations (comprehension), and of onomasiological salience on realizations (production). Clarification will require a range of operationalizations, combining the strengths of corpus data (including frequencies but also recency and cognate effects) with experimental data (keystroke logging, eye-tracking, elicitation and retrospective protocols; e.g. Alves et al, 2010; Serbina & Neumann, 2022).

In this presentation, I highlight a somewhat neglected corpus-based approach to examine the cognitive and linguistic factors underlying translator choices. Malmkjaer (1998) first proposed that, unlike traditional parallel corpora (which contain only one translation for each source text), a multiple translation corpus would offer multiple interpretations and renditions and thus provide a probabilistic overview of translator choices. She pointed out that "where [translators' opinions] differ most is where investigation might prove particularly fruitful." (Malmkjaer,, 1998, p. 6). This multiple translation approach has long remained underexplored for lack of data (but see Castagnoli 2020). However, recent compilation of the Multilingual Student Translation corpus (MUST; Granger & Lefer, 2020), which contains multiple student translations, opens up new avenues to study its potential applications.

My main goal is to demonstrate the feasibility and promise of this approach. As a test case, I present a preliminary study of multiple French translations of English noun sequences (e.g., *disaster relief program coordinator*). Noun sequences represent an extension of the case of noun compounds, whose processing and linguistic representation have been a focus of study in psycholinguistics (Gagné, 2011) and in NLP (Nakov, 2013). In addition, the simple juxtaposition of nouns in these [N+N(+N)…] constructions is efficient and common in English but poses specific difficulties for translation into French. At the formal

level, French tends to use prepositional post-modification [N+Prep+N(+Prep+N)…], which can become problematic for longer sequences (e.g., *disaster relief program coordinator → coordinateur du programme d'aide en cas de catastrophe*). At the semantic level, potentially ambiguous relationships between constituents, left implicit in English, often need to be stated explicitly in French. Both the formal shifts and source item ambiguity are expected to contribute to perceived difficulty, thus often leading to deliberate choices. The present study is based on three English source texts in MUST (in the specialized domain of sustainable finance), from which I extracted 87 noun sequences and 2974 French translations.

I first present a qualitative description of observed differences between the translation solutions. Resulting in part from the multi-word nature of noun sequences, there is an accumulation of variation at three structural levels: the main constituents, their linking prepositions, and surface grammar. I will illustrate how variation at each level can indicate different aspects of translators' interpretations of source ambiguity, in some cases clearly distinguishing multiple interpretations versus multiple realizations.

As suggested by Malmkjaer's quote (above), it is also interesting to analyze the variability of these translations, i.e., the number and distribution of different solutions, and to investigate why some source items give rise to a greater number of solutions. My preliminary quantitative analysis identifies several predictors for high variability: greater noun sequence length, syntactic ambiguity (different plausible dependency relations), and low frequency in English usage (an operationalization of SL salience, obtained from reference corpora for both general and specialized non-translated English). Due to high collinearity, I confirmed the initial multifactorial linear regression model using random forest analysis.

Studying multiple translations and their variability thus appears to be a promising usage-based approach. Both the qualitative and quantitative findings can be related to specific sources of translation difficulty and to the likelihood of deliberate translator choices. In addition, while the multi-word nature of the source items presents a methodological challenge, necessitating analyses of factors and outcomes at different hierarchical levels, these same levels also provide an opportunity for fresh insights.

References

Alves, F., Pagano, A., Neumann, S., Steiner, E., & Hansen-Schirra, S. (2010). Translation units and grammatical shifts: Towards an integration of product-and process-based translation research. In Shreve, G. M., & Angelone, E. (eds). *Translation and cognition* (pp. 109-142). John Benjamins.

Boswijk, V., & Coler, M. (2020). What is salience? *Open Linguistics*, *6*(1), 713-722.

Campbell, S. (2000). Choice Network Analysis in Translation Studies. In: Olahan, M. (ed.) *Intercultural Faultlines. Research Models in Translation Studies 1* (pp. 29-42). St. Jerome.

Castagnoli, S. (2020). Translation choices compared: Investigating variation in a learner translation corpus. In Granger, S. & Lefer, M.-A. (Eds.). *Translating and Comparing Languages: Corpus-based Insights.* Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference (pp. 25-44). Presses Universitaires de Louvain.

De Groot, A. (1997). The cognitive study of translation and interpretation: Three approaches. In Danks, J. H. et al. (eds.). *Cognitive Processes in Translation and Interpreting* (pp. 25-56). Sage Publications.

De Sutter, G., & Lefer, M.-A. (2020). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, *28*(1), 1-23.

Divjak, D. (2019). Frequency in language: Memory, attention and learning. Cambridge University Press.

Gagné, C. L. (2011). Psycholinguistic perspectives. In Lieber, R., & Štekauer, P. (eds.). *The Oxford handbook of compounding* (pp. 255-271). Oxford University Press.

Granger, S., & Lefer, M.-A. (2020). The Multilingual Student Translation corpus: a resource for translation teaching and research. *Language Resources and Evaluation*, *54*(4), 1183-1199.

Halverson, S. L. (2017). Gravitational pull in translation: Testing a revised model. In De Sutter, G., Lefer, M. A., & Delaere, I. (eds.). *Empirical translation studies: New methodological and theoretical traditions* (pp. 9-46). Walter de Gruyter.

Halverson, S. L. (2019). 'Default' translation: A construct for cognitive translation and interpreting studies. *Translation, Cognition & Behavior*, *2*(2), 187-210.

Halverson, S. L. & Kotze, H. (2021). Sociocognitive constructs in Translation and Interpreting Studies (TIS): Do we really need concepts like norms and risk when we have a comprehensive usage-based theory of language? In S. L. Halverson & Á. Marín García (eds). *Contesting Epistemologies in Translation and Interpreting Studies* (pp. 51-79). Routledge.

Heilmann, A., Freiwald, J., Neumann, S., & Miljanović, Z. (2022). Analyzing the effects of entrenched grammatical constructions on translation. *Translation, Cognition & Behavior*, *5*(1), 110-143.

Lefer, M. A., & De Sutter, G. (2022). Using the Gravitational Pull Hypothesis to explain patterns in interpreting and translation: The case of concatenated nouns in mediated European Parliament discourse. In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska & S. Bernardini (Eds.). *Mediated discourse at the European Parliament: empirical investigations* (pp. 133-159). Language Science Press.

Malmkjær, K. (1998). Love thy neighbour: Will parallel corpora endear linguists to translators? *Meta: Translators' Journal*, *43*(4), 534-541.

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3), 291-330.

Serbina, T., & Neumann, S. (2022). Translation Product and Process Data: A Happy Marriage or Worlds Apart? In S. L. Halverson & Á. Marín García (eds.). *Contesting Epistemologies in Translation and Interpreting Studies* (pp. 131-152). Routledge.

Investigating surprisal in simultaneous interpreting

Heike Przybyl[1], Ekaterina Lapshinova-Koltunski[2], Elke Teich[1]

[1]Saarland University, Germany, [2]University of Hildesheim, Germany

KEYWORDS: interpretese, surprisal, cognitive processing, information theory, CBIS

Original speeches differ from simultaneous interpreters' output. The differences described as *interpretese* (Shlesinger, 2009) are assumed to originate in the specific production constraints, linked to particular cognitive processing and production effort in simultaneous interpreting. Interpreters tend to use more verbal output whereas comparable originals are found to be more nominal (Przybyl et al., 2022a), interpreters produce more standardised, formulaic output than impromptu source speakers (Kajzer-Wietrzny and Grabowski, 2021) and are found to use generic nouns more frequently than original speakers (Bizzoni et al., 2021).

The current study tries to link *interpretese* to cognitive processing differences by using surprisal as probabilistic measure based on Information Theory (Shannon, 1948), approximating cognitive processing in simultaneous interpreting and original production.

Generally, texts are assumed to vary in the level of information content they convey. It is assumed that information conveyed by texts is distributed evenly over a message, aiming to avoid information peaks and troughs (Levy and Jaeger, 2006). Informativity above or below expected rates may lead to processing difficulties by the recipient of the message (Engelhardt et al., 2011; Kravtchenko and Demberg, 2015). Degaetano-Ortlieb and Teich (2022) assume that linguistic options available to the encoder are used to modulate the level of information encoded in the message to achieve optimal encoding. This can be measured with surprisal as word-based measure of cognitive effort (Hale, 2001). Words that are very unpredictable are high in surprisal and therefore require high cognitive processing effort.

Few studies in the translation field have used information theoretic measures to compare original texts with translations (Bizzoni and Lapshinova-Koltunski, 2021; Teich et al., 2020; Martinez and Teich, 2017; Rubino et al., 2016). Lapshinova-Koltunski et al. (2022) investigate surprisal of discourse connectives in original speeches and simultaneous interpreting. The current study looks at lexical words and uses *surprisal* or *unpredictability* in context to approximate cognitive processing by the producers, in our case original speakers and simultaneous interpreters. We hypothesise that – due to the particular production constraints in simultaneous interpreting – interpreters will use more expected lexical words than comparable original speakers in the same language.

The study uses English and German original and interpreted speeches of the European Parliament, a subset of the EPIC-UdS corpus (Przybyl et al., 2022b). The corpus consists of transcripts of European Parliament speeches, including information on delivery mode of the original speeches (whether original speeches were prepared and read out, impromptu or a mixed delivery mode). Furthermore, the corpus data include surprisal annotation (cf. Equation 1), measured as information content in number of bits, calculated as negative log base 2 probability of a given occurrence of a unit (e.g. word) in a given context (cf. Equation 1). Context is defined by the preceding words (here: 3 words) for the current study.

$$S(word) = -log2p(word|context) \qquad (1)$$

Distinctive part-of-speech for interpreting and original speech respectively (Przybyl et al., 2022a) are investigated by (1) comparing average surprisal for verbal and nominal verb classes in a comparable corpus analysis, also considering potential effects resulting from different delivery modes and (2) looking at potential triggers and translations of particularly high and low surprisal lexical items in a parallel analysis. More expected words are easier to produce and therefore we assume that interpreters use more expected words in context than originals speakers.

Our results show a general tendency for more expected words used in simultaneous interpreting. On average, interpreters produce lower surprisal words than original speakers. The trend is confirmed for verbs in the comparable and parallel analysis and can also be found for comparable nouns in the English

dataset. Delivery mode generally confirms the expected trend of simultaneous interpreting using lower surprisal words than impromptu original speeches, however this is only statistically significant for the comparable English dataset and not for German.

This means that interpreters as recipients of original speech need to process more unexpected words in context than the recipients of interpreted speech. By producing lower surprisal output, interpreters facilitate processing for the listeners. Whether this is audience design (intended by the interpreter to reduce processing costs for the listener) or resulting from their own processing constraints cannot be answered by means of this analysis. However, further results from the parallel analysis such as high surprisal nouns in interpreting being triggered by cognates of the source speech point to a producer-oriented mechanism.

Interpreters produce more expected output overall. However, results for the languages studied differ, showing a clear trend for lower surprisal words overall in English and some exceptions of this trend in German. Although the results of the current study are very promising, the application of the current methodology also showed weaknesses: the dataset used for modelling is very small and therefore hapax legomena occur frequently, distorting the current results especially for low frequency investigations. We intend to tackle this issue in a follow up study by using a larger general language model as high quality in-domain transcribed data, particularly for the languages studied, is rare.

References

Bizzoni, Y. and Lapshinova-Koltunski, E. (2021). Measuring translationese across levels of expertise: Are professionals more surprising than students? In Proceedings of the 23rd NoDaLiDa, pages 53–63, Online. ACL.

Bizzoni, Y., Przybyl, H., and Teich, E. (2021). Cutting semantic corners? patterns of lexical simplification in interpreting vs. translation. In Castagnoli, S., Bernardini, S., Ferraresi, A., and Milicevic Petrovic, M., editors, Using Corpora in Contrastive and Translation Studies Conference (6th Edition).

Degaetano-Ortlieb, S. and Teich, E. (2022). Toward an optimal code for communication: The case of scientific English. Corpus Linguistics and Linguistic Theory, 18(1):175–207.

Engelhardt, P. E., Barıs Demiral, and Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An erp study. Brain and Cognition, 77(2):304–314. Special Section: Aggression and peer victimization: Genetic, neurophysiological, and neuroendocrine considerations.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics, pages 159–166, Stroudsburg, PA. Association for Computational Linguistics.

Kajzer-Wietrzny, M. and Grabowski, (2021). Formulaicity in constrained communication an intermodal approach. MonTI. Monografias de Traduccion e Interpretacion.

Kravtchenko, E. and Demberg, V. (2015). Semantically underinformative utterances trigger pragmatic inferences. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., and Maglio, P. P., editors, CogSci. cognitivesciencesociety.org.

Lapshinova-Koltunski, E., Pollkläsener, C., and Przybyl, H. (2022). Exploring explicitation and implicitation in parallel interpreting and translation corpora. The Prague Bulletin of Mathematical Linguistics, 119:5–22.

Levy, R. and Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. Advances in Neural Information Processing Systems, 19:849–856.

Martinez, J. M. M. and Teich, E. (2017). Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In Cercel, L., Agnetta, M., and Lozano, M. T. A., editors, Kreativität und Hermeneutik in der Translation. Narr Francke Attempto Verlag.

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

Przybyl, H., Lapshinova-Koltunski, E., Menzel, K., Fischer, S., and Teich, E. (2022b). EPIC UdS - creation and applications of a simultaneous interpreting corpus. In Proceedings of the Language Resources and Evaluation Conference, pages 1193–1200, Marseille, France. European Language Resources Association.

Rubino, R., Lapshinova-Koltunski, E., and van Genabith, J. (2016). Information density and quality estimation features as translationese indicators for human translation classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 960–970, San Diego, California. Association for Computational Linguistics.

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27:379–423, 623–656.

Shlesinger, M. (2009). Towards a definition of interpretese: An intermodal, corpus-based study. In Chesterman, A. and Gerzymisch-Arbogast, H., editors, Efforts and Models in Interpreting and Translation Research: A tribute to Daniel Gile, pages 237–253.

Teich, E., Martinez Martinez, J., and Karakanta, A. (2020). Translation, information theory and cognition. In Alves, F. and Jakobsen, A. L., editors, The Routledge Handbook of Translation and Cognition, chapter 20. Routledge, London.

Recipe for a bilingual Portuguese-English gastronomic glossary

Rozane Rodrigues Rebechi, Nathália Oliva Marcon, Patrícia Helena Freitag

Universidade Federal do Rio Grande do Sul, Brazil

KEYWORDS: Restaurant review, genre, functional translation, corpus linguistics, conventionality

A restaurant review serves as a means for professional critics to share their impressions on food, ambiance, and service (Blank, 2007), employing colorful and engaging language that resembles a literary style (Pudlowski, 2012). Although restaurant reviews share certain genre-specific similarities, a closer examination reveals that the objective is achieved through diverse strategies that vary across languages and cultures. This paper explores how a corpus-based methodology can inform us about linguistic and cultural differences that should be considered when translating or writing restaurant reviews for two distinct target audiences: Brazilians and North Americans. In order to produce texts that effectively resonate with the target readers, translators must be aware of the conventions within this domain in both languages and cultures, enabling them to decide which aspects should be maintained, adapted, or omitted (Rebechi et al., 2021). For the language pair of Brazilian Portuguese and North American English, the available reference material to assist translators in the challenging task of rendering culinary terminology is still limited. To contribute to bridging this gap, we developed a research project named "Culinária para Fins Acadêmicos: compilação de um corpus de textos culinários com foco na tradução" (Culinary for Academic Purposes: the compilation of a corpus of culinary texts for translation). The project was especially concerned with compiling entries and searching for their functional equivalents, with the ultimate goal of creating a freely accessible online and bidirectional English-Portuguese glossary of terms and collocations characteristic of restaurant reviews, Dicionário Gastronômico (2020) (Gastronomic Dictionary).

By adopting corpus linguistics as our methodology, we conducted both quantitative and qualitative analyses on a comparable corpus of restaurant reviews published in the United States and Brazil from 2016 to 2019. The corpus, comprising 441 reviews in English and 1,266 reviews in Portuguese, results in approximately 420,000 tokens in each language. Through manual investigation of simple and compound keywords retrieved using Sketch Engine (Kilgarriff et al., 2014), we concluded that corpora not only assist translators in interpreting source language texts but also shed light on the usage and meaning of words and expressions specific to the genre, aiding professionals in finding translation solutions. This is particularly true when translators aim to produce functional translations (Nord, 2006), which go beyond mere terminological equivalence and also consider the genre's specificities in both languages and cultures. Our analysis revealed several noteworthy findings. Firstly, we observed that English restaurant reviews are, on average, three times longer than their Brazilian counterparts, resulting in a larger number of keywords. Upon examining these keywords, we concluded that English texts not only prioritize the product itself, such as the evaluated food or drink, but also highlight the process of preparation, including reference to the utensils used. As a result, we encountered many terms in the English corpus for which we could not find equivalents in Portuguese. One example is the keyword "cast-iron skillet," whose prima facie equivalent – "*frigideira de ferro*" - was not found in the Portuguese subcorpus.

Moreover, we also discovered that Brazilian critics tend to employ less specialized terminology when referring to techniques and utensils. For instance, in relation to cuts of meat and vegetables, the English corpus yielded various results such as "minced," "(finely) chopped," "cubed," "diced," and "julienned," indicating precision in the processes. In contrast, the preferred strategy in Portuguese seems to be the use of diminutives to indicate the difference between cuts, such as "*(bem) picadinho*" encompassing the first four cuts in English and "*cortado fininho*" for the last one.

In our analysis, we identified additional dissimilarities between the corpora of restaurant reviews. These include the use of metaphors that stem from distinctive cultural phenomena, such as references to carnival

in the Portuguese corpus and baseball in the English corpus. Furthermore, we found quantitative and qualitative differences in the usage of adjectives and adverbs between Portuguese and English texts. The English corpus exhibited a greater variety and frequency of adjective types, as well as a larger proportion of adverbs ending in "-ly" compared to their Brazilian counterparts, which typically end in "*-mente*." Our findings led us to conclude that restaurant reviews constitute a genre rich in culture-specific terms that not only lack direct counterparts in the target language but are also closely tied to cultural aspects that communicate beyond culinary matters alone. Some of these distinctions, pertaining to the Brazilian Portuguese/North American English language pair, have been previously identified and discussed in other research that analyzed texts laden with cultural significance, such as obituaries (Rebechi & Silva, 2018) and hotel sites (Fuchs, 2018). By highlighting the discrepancies between the keyword lists, we demonstrate that rendering a source text term with its prima facie equivalent, even when a suitable equivalent exists, may not always be the optimal or appropriate choice for producing a conventional text (Tagnin, 2013). The methodology employed in this study underscores the importance of keywords in both languages for identifying potential equivalents, as well as the significance of the absence of direct correspondence between terms in revealing differences that should be considered during the translation of this genre. Finally, it is worth noting that since our corpus is relatively small, expanding it for future research would be advisable.

References

Blank, G. (2007). Critics, ratings, and society: The sociology of reviews. Rowman & Littlefield Publishers.

Dicionário Gastronômico. (2020). Porto Alegre: Universidade Federal do Rio Grande do Sul. Retrieved from https://www.ufrgs.br/dicionariogastronomico/

Fuchs, S. L. M. (2018). Glossário bilíngue de colocações de hotelaria: um modelo à luz da Linguística de Corpus. (Doctoral dissertation, Universidade de São Paulo, São Paulo).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M. š., Kovář, V. ě., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1(1), 7-36.

Nord, C. (2006). Loyalty and fidelity in specialized translation. Confluências: Revista de Tradução Científica e Técnica, 29-41.

Pudlowski, G. (2012). Para que serve um crítico gastronômico? Tapioca.

Rebechi, R. R., Nunes, R. R., Munhoz, L. R., & Marcon, N. O. (2021). Restaurant Reviews in Brazil and the USA: A Feast of Cultural Differences and Their Impact on Translation. MUTATIS MUTANDIS (MEDELLIN. 2008), 14, 372-396.

Rebechi, R. R., & Silva, M. M. (2018). Obituaries in translation: A corpus-based study. Cadernos de Tradução (UFSC), 38, 298-318.

Tagnin, S. E. O. (2013). O jeito que a gente diz. Disal Editora.

Do translations from comparatively high-status source languages exhibit more lexical interference? A multilingual corpus-based analysis

Matt Riemland

Dublin City University, Ireland

KEYWORDS: language status, loanwords, borrowings, interference, corpus-based translation studies, comparable corpora

This presentation constitutes a substantial evidentiary and analytical expansion of the underlying study's preliminary results, which were presented at the Translation in Transition VI Conference at Charles University in September 2022. The study's research question asks whether translations from comparatively higher-status source languages (SLs) tend to contain more evidence of lexical interference than those from comparatively lower-status SLs. In translation studies, interference refers to the reproduction of typical source-language features in the target text. The positive correlation between a translated text's comparatively higher SL language status – relative to that of its target language (TL) – and its measurable traces of interference has been hypothesized since the early days of corpus-based translation studies (see Baker, 1993, p. 183; Toury, 1995/2012, p. 314). Despite previous research interrogating this phenomenon for individual language pairs (see Mauranen, 2004; Becher et al., 2009; Evert and Neumann, 2017), a systematic study involving a range of languages varying in status has been lacking. This study aims to address this gap.

While language prestige denotes individuals' personal attitudes towards languages, language status may be distinguished as the relatively apparent social positioning of languages with respect to other languages in their proximity (Edwards, 1996, p. 703). In order to assess language status, the study adapts the sociolinguistic framework put forward by Lewis and Simons (2010), which arranges languages into hierarchical categories according to their established use in various social domains (e.g. international commerce, education, religious rites). This study focuses on seven European languages (presented from the highest to lowest status): English, French, German, Italian, Swedish, Croatian, and Irish. It applies comparable corpus methodology to a multilingual corpus (roughly 26 million tokens) of translated and non-translated literary prose published in the late 19th and early 20th centuries. Each translation's source language is one of the study's other selected languages. Each selected language also has its own comparable subcorpus of texts originally produced in that language. Language status discrepancies between source languages and target languages are determined according to the difference between each language's "score" within the list; English has the highest score (7), and Irish has the lowest score (1). Lexical interference is operationalized here as the relative frequency of *translational* loanwords (TWLs) – a newly coined term. TLWs are loanwords that are reasonably attributable to the target text's production process; this definition excludes loanwords stemming from prior contact between the source and target languages, where the term has been previously adopted into the target language's standard lexicon.

The study determines likely TLW candidates by comparing words' relative frequencies across the relevant corpora. The first list of likely TLW candidates is comprised of those words whose absolute frequency in the SL comparable corpus is greater than their absolute frequency in the translation corpus, and which do not appear in the TL comparable corpus. In order to account for TLWs among interlingual homonyms, the second list of likely TLW candidates contains word forms appearing in both the SL and TL comparable corpora as well as the translation corpus. Concordances for each possible TLW candidate are then manually reviewed in order to identify TLWs. Monolingual dictionaries are also consulted in order to determine whether the TLW candidate had already been established in the TL lexicon, in which case the word is disqualified as a TLW. Collocations of up to 5 grams are counted as single TLW units. Lexical items such as proper nouns, book titles, quoted poems or songs, currencies, orthographically adapted source-language words, and source-language words already established in the target language's lexicon are not counted as TLWs. TLW candidates originating from languages other than a target text's respective

SL – such as English loanwords in German translations of French source texts – do not qualify as TLWs, as they do not reflect a form of lexical interference that may be narrowly attributed to the target text's local translation process. The study quantifies lexical interference as the relative frequency of TLWs in translated texts. Language status discrepancies therefore constitute the explanatory variable and the relative frequencies of TLWs constitute the response variable.

Whereas the past demonstration of this study's preliminary results attempted to approximate a linear correlation between language status and lexical interference, this updated presentation recharacterizes and hypothesizes the relationship between these two variables as an increasing monotonic trend. Monotonicity denotes a function's consistency in increasing or decreasing, regardless of the rate at which it increases or decreases. The study's new statistical approach – Spearman's rank correlation coefficient – produces a more theoretically sound data analysis, given the nature of the explanatory language status variable (i.e. a ranked variable, where the distance between rankings is ambiguous). The results presented here are also based upon a significantly cleaner and expanded data set than that of the preliminary results. The study determines a weakly decreasing monotonic trend between language status and TLWs, providing evidence that translations from comparatively higher-status languages *do not* exhibit more traces of lexical interference. Consequently, the presentation provides alternative vantage points for the data, emphasizing the possible influence of imbalances in the authors, translators, language pairs, and texts included in the corpus. It also posits register as a potential confounding variable. This study may be replicated in domains beyond the literary genre; future studies are also suggested to include language status as a variable in multivariate analyses.

References

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In honour of John Sinclair* (pp. 233–250). John Benjamins Publishing Company.

Becher, V., House, J., & Kranich, S. (2009). Convergence and divergence of communicative norms through language contact in translation. In K. Braunmüller & J. House (Eds.), *Convergence and divergence in language contact situations* (pp. 125–152). John Benjamins.

Edwards, J. (1996). Language, prestige and stigma. *Kontaktlinguistik*, 703–708.

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In *Empirical Translation Studies: New Methodological and Theoretical Traditions* (pp. 47–80). De Gruyter Mouton. https://www.degruyter.com/document/doi/10.1515/9783110459586-003/html

Lewis, M., & Simons, G. (2010). Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, *55*. https://doi.org/10.1017/CBO9780511783364.003

Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals: Do they exist?* John Benjamins Publishing Company.

Toury, G. (1995). *Descriptive Translation Studies - and Beyond: Revised Edition* (2nd ed.). John Benjamins Publishing Company.

Playing hide-and-seek with linguistic categories in a parallel corpus: Universal Dependencies at work

Alexandr Rosen

Charles University, Czechia

KEYWORDS: parallel corpus, morphosyntactic annotation, Universal Dependencies, politeness, finiteness

In this contribution we try to identify possibilities and limits of language-universal linguistic annotation applied to a parallel corpus. We are concerned with aspects useful from the perspective of contrastive and translation studies and focus on categories which the annotation does not make explicit.

Standard linguistic categories do not mean the same across languages. Instead of language-universal categories, Haspelmath (2010) proposes *comparative concepts*, requiring potentially non-trivial mappings to language-specific categories. For Ramscar & Port (2019), all categories vary with context. The issue comes up in the annotation of corpora, especially parallel corpora. Nowadays the prevailing strategy is to adopt for all languages a uniform annotation standard, such as *Universal Dependencies* (henceforth *UD*, de Marneffe et al. 2021).[1] Although *UD* suffers both from the theoretical problems inherent to a language-universal taxonomy[2] and from the practical problems of applying the guidelines consistently across more than 100 languages, it has been steadily gaining ground.[3] Recently, a release of *InterCorp*, a large multilingual parallel corpus,[4] was published with two annotation options: language-specific tagsets and *UD*, produced by the *UDPipe* tool (Straka 2018).[5] Future releases of *InterCorp* are due with the *UD* annotation as the only option. Since *InterCorp* is used mainly via *KonText*,[6] an on-line token-based concordancer, the *UDPipe* output (in the CONLL-U format) is modified to allow easy access to categories and syntactic structure in queries and statistics.[7]

It seems that at least for some tasks involving cross-lingual comparison, the drawbacks of language-universal categorization are outweighed by the benefits of a uniform annotation scheme and the fact that syntactic annotation, available in *UD*, is less sensitive to typological differences. Yet our first goal is to verify an even more ambitious claim that the *UD* annotation can be used to detect categories which are not labelled explicitly.

Our second goal takes advantage of the assumed meaning equivalence between parallel texts in the corpus, again with the aim to remedy some gaps in existing annotation. While available morphosyntactic annotation schemes refrain from specifying categories which require decisions involving a larger context or discourse knowledge, such cases are often language specific. Thus, a *UD*-annotated parallel corpus can be used (i) to resolve ambiguities using cross-lingual information in parallel concordances, (ii) to study contrasts in usage, or (iii) to investigate translation strategies, depending on the source and target languages. In other words, our second goal is to see how far we can proceed in (i) – (iii).

We focus on two categories of Czech, German and Polish verbs or verbal constructions: politeness and finiteness, as expressed by morphological or syntactic patterns, yet – at least in UD – not annotated explicitly or unambiguously. Looking for cues in the annotation available in the Czech texts or in the

---

[1] https://universaldependencies.org

[2] See, e.g., Croft et al. (2017), Cimiano et al. (2020: 137–140), Kutuzov et al. (2016). For an alternative approach to representing syntactic relations see Gerdes et al. (2018).

[3] Even in fields such as translation studies or language typology, cf. Lapshinova-Koltunski (2021), Poirier (2021), or Levshina (2022).

[4] https://wiki.korpus.cz/doku.php/en:cnk:intercorp

[5] https://ufal.mff.cuni.cz/udpipe/2

[6] https://www.korpus.cz/kontext/query?corpname=intercorp_v13ud_en&align=intercorp_v13ud_pl

[7] For details, see https://wiki.korpus.cz/doku.php/en:pojmy:ud.

parallel German and/or Polish texts, the two case studies are mainly concerned with Czech, where information gaps in these two categories seem to be most obvious. The results will be presented together with an insight into (i) the crucial properties of the *UD*-based annotation and (ii) the parallel search interface of *InterCorp*.

In the Czech polite forms, the finite verbal form marked for person is plural, agreeing with the (usually dropped) pronominal 2nd person plural subject (like in French). Without any participial or adjectival form in the predicate, which is singular when a single person is addressed politely, the form is ambiguous between plural and polite singular. On the other hand, the Polish and – to some extent – German polite forms are unambiguous. Moreover, the comparison reveals interesting differences in usage, reflecting additional pragmatic and sociolinguistic phenomena.

The second case study concerns the contrast between finite (personal) and non-finite forms: infinitives, converbs and deverbal nouns or participles. While some Czech finite forms are annotated as participles, other participles or deverbal nouns are annotated as adjectives or plain nouns. Many of these cases could be resolved using parallel concordances. Again, the study also compares the usage of finite and non-finite forms in the three languages.

To investigate the parallel Czech, German and Polish data, two subcorpora are extracted from the *UD*-annotated release of *InterCorp*, both restricted to fiction. The larger subcorpus includes about 15 mil. tokens represented by a single set of texts in each language, irrespective of whether and from which language the text is translated, i.e., for some texts, the subcorpus includes only translations from another language. This subcorpus is used to count the relative frequencies of usage of the categories and the corresponding patterns in the three languages.

The smaller subcorpus, about 4 mil. tokens in a language, includes only non-translated texts in one of the three languages and their translations in the other two. The data derived from this subcorpus distinguish translated and non-translated texts and are used to bring the dimension of translation into the results. Although none of the two subcorpora is balanced in terms of the share of non-translated texts, it only affects the significance of the translation-related results for Polish.[8]

To identify all patterns of the observed phenomena, syntactic annotation is needed: some of the structures are analytical verb forms, represented in *UD* as syntactic trees, in other forms it is important to identify the subject (e.g., *Sie* in German or *Państwo*, *Pani*, etc. in Polish). As the first step, the unambiguous patterns in all the three languages are detected using monolingual queries. Next, querying the Czech part of the subcorpus, other potential positives are identified. Finally, the Czech candidates for the given categorial status are filtered using equivalents in German or Polish or both. Due to the missing word-to-word alignment in the annotation and the corresponding functionality in the concordancer, some manual intervention is needed in this step to exclude cases of accidental occurrences of target patterns. Otherwise, the quantitative parts of both studies use the statistics provided by the concordance directly.

The results show a slight predominance of plain forms over polite forms in Czech (62%) and German (54%), contrasted with an overwhelmingly high share of plain forms in Polish (94%). Finite verbal forms are more frequent in all the three languages, with Polish at 83%, Czech at 46% and German at 35%. These results are confirmed and some hypothesis about their causes tested in the balanced subcorpus and by manual analysis of parallel concordances.

References

Bisiada, M., editor (2021). *Empirical studies in translation and discourse*. Number 14 in Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

[8] In the subcorpus of the same texts in the three languages, Polish has a lower amount of non-translated texts than Czech or German.

Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer.

Croft, W., Nordquist, D., Looney, K., & Regan, M. (2017). Linguistic typology meets Universal Dependencies. In Dickinson, M., Hajič, J., Kübler, S., & Przepiórkowski, A., editors, *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75. Indiana University, Bloomington, Bloomington, IN, USA.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*, Brussels, Belgium.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Kutuzov, A., Velldal, E., & Øvrelid, L. (2016). Redefining part-of-speech classes with distributional semantic models. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 115–125, Berlin, Germany. Association for Computational Linguistics.

Lapshinova-Koltunski, E. (2021). Analysing the dimension of mode in translation. In Bisiada (2021: 221–241).

Levshina, N. (2022). Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology* 26(1):129–160.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Poirier, E. A. (2021). What can Euclidean distance do for translation evaluations? In Bisiada (2021: 165–198).

Ramscar, M. & Port, R. (2019). Chapter 4: Categorization (without categories). In E. Dąbrowska & D. Divjak (Ed.), *Cognitive Linguistics – Foundations of Language* (pp. 87–114). Berlin, Boston: De Gruyter Mouton.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Exploring the interface between corpus and statistical modelling: an analysis of lexical variation in Islamic legal discourse

Rana Roshdy

Dublin City University, Ireland

KEYWORDS: corpus linguistics, genre variation, profile-based correspondence analysis, logistic regression, Islamic law and finance

Aims

This research sets out to investigate how 'culture-specific' or 'signature' concepts are rendered in English-language discourse on Islamic, or 'sharia' law, which has Arabic roots. A growing body of literature has investigated Islamic law from a technical perspective. However, from the perspective of linguistics and translation studies, little attention has been paid to the lexicon that makes up this specialised discourse. Much of the commentary has so far been prescriptive, with limited empirical evidence. This study aims to bridge this gap by exploring how 'culturalese', i.e., ostensive cultural discourse (Roshdy, 2023), travels through language, as evidenced in a self-built monolingual English corpus on Islamic financial and family law matters. It will investigate non-canonical forms of translation embedded in the discourse of Islamic law.

The English-language discourse on Islamic law, both translated and nontranslated, offers mediation of an Arabic-origin culture, therefore embodying the paradigm of 'cultural translation', which refers to the transfer of concepts or ideas from one language or culture into another. Exhibiting translation as a radical force of resistance, the key objective of cultural translation is to create a hybrid space for the heritage of less dominant cultures to be expressed and represented in the dominant languages (Pym, 2014).

Corpus information

This study is based on the Islamic Law Corpus (ICL), a specialised monolingual corpus comprising over 9 million words and is divided thematically into two subcorpora: Islamic finance and family law. The corpus is considered multi-genre since each subcorpus is composed of five genres. Each genre either has a performative or a non-performative function based on whether it is intended for legal application or for information purposes only. The performative category comprises applied genres that have a binding force, including 'law and policy-making instruments' (e.g., codes, standards, policy documents) and 'other instruments' (e.g., contracts, agreements, application forms), while the non-performative category covers purely informative genres (e.g., books, academic articles, grey literature). This design will identify whether culture-specific concepts are expressed differently across diverse contexts or genres.

Research questions

The questions driving this research are varied in nature. The primary question of this research is descriptive, which generally seeks to describe a phenomenon based on the extracted data:

- RQ1: What are the most frequent strategies used to translate culture-specific concepts from Islamic law into English?

The secondary questions are explorative, seeking to explore three predicator variables: function, genre, and subject field:

- RQ2: Is lexical variation influenced by the function of the legal texts in question (performative versus non-performative)?
- RQ3: Are there inter-genre differences within each of these categories? (e.g., books vs academic articles in the non-performative category; policy-and law-making instruments vs other instruments in the performative category)

- RQ4: Are there inter-subject field differences within each of these categories? (finance vs family law)

Multi-methodological design: corpus, statistical modelling, emancipatory theories

The research adopts a mixed methods design. The study first quantifies the different linguistic strategies used to render sharia-based concepts in English, with reference to Šarčević's (1985) typology, in order to explore 'translation' norms based on linguistic frequency in the corpus. This quantitative analysis employs statistical logistic regression using MATLAB programming software. The findings are then interpreted qualitatively in light of emancipatory postcolonial translation agendas, which aim to preserve intangible cultural heritage and promote the representation of minoritised groups.

Using Sketch Engine software, the analysis will start with a bottom-up query to build a profile set through a 'profile-based correspondence analysis' (Delaere & De Sutter, 2017), which considers the probability of lexical variation in expressing a conceptual category. Firstly, through the 'keywords' function, Arabic loanwords that feature in the English-language discourse on Islamic law will be extracted. Secondly, 'collocations' and 'concordance' analysis will be employed to identify loanwords that have an 'endogenous' counterpart lexeme, meaning a close English equivalent or a synonymous naming alternative (Delaere & De Sutter, 2017). To establish the data sets, the study will focus on 20 lexical profiles (20 Islamic concepts or conceptual systems: 10 from finance and 10 from family law); each profile will consist of the Arabic loanword and all corresponding endogenous variants for the same conceptual category. Thirdly, a concordance analysis of the selected profiles will investigate the different translation strategies. The concordance analysis will involve annotating a random sample of 8000 concordance lines covering the 20 concepts. For each concept, 200 instances of the loanword versus 200 instances of the identified endogenous labels will be annotated.

Finally, logistic regression will be implemented to measure the influence of the explanatory variables, 'genre', 'legal function' (i.e., performative versus non-performative function), and 'subject field' on the choice between an Arabic loanword and an endogenous English lexeme. This regression model involves searching for the variants of each 'profile' in each of the different genres and then expressing the probability of each lexical variant occurring in a particular genre, legal function, or subject field as a percentage. The efficacy of the logistic regression analysis is due to its ability to determine the simultaneous impact of different predictor variables on a response variable.

Results

The research finds that the English-language discourse on Islamic law is characterized by linguistic borrowing and glossing, implying an ideologically driven variety of English that can be usefully labelled as a kind of 'Islamgish' (blending 'Islamic' and 'English') aimed at retaining symbols of linguistic hybridity. The regression analysis confirms the influence of contextual factors (genre, function, and subject field) on the use of an Arabic loanword versus English alternatives. A comparative evaluation of the results suggests that Islamic finance demonstrates a higher scale of linguistic borrowing due to the greater cultural specificity of its terminology, compared to the other branch of Islamic family law, thus ascertaining the subject field as a variable impacting lexical variation.

References

Pym, A. (2014). *Exploring translation theories*. (2nd ed.). Routledge.

Roshdy, R. (2023). Translating Islamic law: the postcolonial quest for minority representation. [PhD Thesis. Dublin City University]. DORAS.

The MathWorks, Inc. (2023). *MATLAB* [Software]. Available at: https://www.mathworks.com/products/matlab.html (Accessed: 9 January 2023).

Lexical Computing CZ s.r.o. (2003). *Sketch Engine* [Software]. Available at: https://www.sketchengine.eu/ (Accessed: 9 January 2023).

Delaere, I. & De Sutter, G. (2017). Variability of English loanword use in Belgian Dutch translations: measuring the effect of source language and register. In G.De Sutter, M. Lefer & I. Delaere, (Eds.) *Empirical translation studies: new methodological and theoretical traditions* (pp. 81-112). De Gruyter Mouton.

The processing of numbers in simultaneous interpreting: The effect of directionality, speech rate, and mode of delivery

Paulina Rozkrut

Adam Mickiewicz University, Poznań, Poland

KEYWORDS: Corpus-based Interpreting Studies, simultaneous interpreting, number interpreting, interpreting accuracy, problem triggers

Simultaneous interpreting tends to be seen as one of the most extreme and cognitively demanding tasks that can be performed by human beings (Obler, 2012; Seeber, 2021). Processing information in two different languages concurrently and having to multitask under time pressure are among the main reasons why interpreters experience a significant level of stress. The difficulty of interpreting becomes even more pronounced when "problem triggers" (Gile, 2009) such as numbers come at play. So far, research has shown that numerical expressions pose a great challenge for interpreters. However, it must be noted that the vast majority of studies have investigated this issue in the experimental setting and with students as participants (e.g., Braun & Clarici, 1996; Mazza, 2001; Pinochi, 2009; Desmet et al., 2018; Frittella, 2019), with only a few works exploring how professional interpreters deal with numbers in the natural interpreting context (e.g., Collard & Defrancq, 2017; Kajzer-Wietrzny et al., 2021).

In view of the scarcity of research using naturalistic data, the aim of this study, which was conducted as part of an MA thesis, was to explore how professional simultaneous interpreters cope with numbers, and what factors affect the accuracy of their performance. The following research questions were addressed:

1. Is the accuracy of number interpreting affected by such factors as directionality, speaker's speech rate, or mode of delivery of the source text?
2. Which categories of numbers are most and least likely to be interpreted correctly?
3. Which category of incorrect renditions of numbers is most frequent?

Owing to its advantages, a naturalistic, corpus-based research approach was adopted to answer the questions. In contrast to experimental data, naturalistic data come from actual conferences that have been interpreted; consequently, they present a more reliable view of the work of interpreters, who have not been constrained by the research design (Shlesinger, 1998; Setton, 2011). Additionally, corpora provide the opportunity to examine large amounts of data and thus observe "trends, patterns, concordances, common and uncommon phenomena" (Bendazzoli, 2018, p. 7) that may go unnoticed in experiments.

The corpus used in the study was PINC: Polish Interpreting Corpus (Chmiel et al., 2022). Similarly to other interpreting corpora based on the European Parliament sittings (see Monti et al., 2005; Kajzer-Wietrzny, 2012; Defrancq et al., 2015; Bernardini et al., 2018), PINC is composed of plenary speeches given by MEPs and their simultaneous interpretations. On the whole, the corpus includes four subcorpora of (1) Polish source speeches (40,168 tokens), (2) English interpretations (41,283 tokens), (3) English source speeches (54,089 tokens), and (4) Polish interpretations (36,649 tokens) that vary, among others, in terms of length, delivery mode, topics, speakers, and interpreters (Chmiel et al., 2022).

The data analysis process was divided into two stages. During the first stage, a list of number categories and rendition categories was compiled on the basis of the study by Kajzer-Wietrzny et al. (2021). During the second stage, each source and target speech from the PINC corpus was examined in search of numerical expressions. Whenever a number was identified, it was labeled with one of the adopted categories of numbers (in the source speech) and renditions (in the target speech). Also, speech rate (measured in the number of syllables uttered per minute) and the mode of delivery (read-out vs. impromptu) were aligned with each source text. Ultimately, a total of 986 numbers were subject to statistical analyses (linear mixed models) performed in R Studio.

The study found that number rendition is largely determined by the interpreting direction. Specifically, it was revealed that the accuracy of number interpreting was higher for interpretation from the mother tongue into the foreign language (69% of accurate renditions) than vice versa (61% of accurate renditions), corroborating the results reported by Braun & Clarici (1996) and Kajzer-Wietrzny et al. (2021). The speech rate, in turn, was not observed to affect the interpreting accuracy, thus contradicting the findings of experimental research by Korpal & Stachowiak-Szymczak (2020) but confirming those obtained by Kajzer-Wietrzny et al. (2021), who analyzed naturalistic data. Such a variance may indicate that the help of a boothmate alleviates the difficulty of number interpreting under a faster speech rate, as opposed to the necessity of working in isolation during experiments. As for the effect of delivery mode, the analysis showed that impromptu speeches were interpreted more accurately than those that had been read-out (70% vs. 60% of accurate renditions), which points to easier processing of numbers in spontaneous presentations. Given the scarcity of research on this subject, it may be worthwhile for future studies to further analyze the finding and extend it to other language pairs.

The results also included descriptive statistics for number categories that were most and least prone to accurate interpretation as well as the categories of inaccurate renditions that were most common. While the distribution of the most and least accurately interpreted number categories differed across the two directions, the analysis of inaccuracies revealed that omission was noticeably the most frequent inaccurate rendition for interpreting into both the mother tongue and the foreign language and thus corroborated the results of previous research by Braun & Clarici (1996), Mazza (2001), Pinochi (2009), Desmet et al. (2018), and Kajzer-Wietrzny et al. (2021).

This work contributes to our knowledge about processing numbers in the natural interpreting environment. As such, it may be of interest not only to practitioners but also to lecturers and trainees, drawing their attention to the factors that may affect number rendition. Since some of such factors are beyond the control of interpreters (speech rate, delivery mode, interpreting direction), teaching students how to deal with these challenges is crucial. At the same time, the work is a contribution to the debate on directionality, which for a long time has aimed to determine the superiority of either comprehension or production in the native language. As indicated by the findings, the superiority of one interpreting direction over the other may depend not only on the status of the interpreter's working language but also on the presence of problem triggers such as numbers.

References

Bendazzoli, C. (2018). Corpus-based interpreting studies: Past, present and future developments of a (wired) cottage industry. In M. Russo, C. Bendazzoli, & B. Defrancq (Eds.), *Making way in Corpus-based Interpreting Studies* (pp. 1–19). Springer. https://doi.org/10.1007/978-981-10-6199-8

Bernardini, S., Ferraresi, A., Russo, M., Collard, C., & Defrancq, B. (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In M. Russo, C. Bendazzoli, & B. Defrancq (Eds.), *Making way in corpus-based interpreting studies* (pp. 21–42). Springer. https://doi.org/10.1007/978-981-10-6199-8

Braun, S., & Clarici, A. (1996). Inaccuracy for numerals in simultaneous interpretation. *The Interpreters' Newsletter*, 7, 85–102.

Chmiel, A., Kajzer-Wietrzny, M., Koržinek, D., Janikowski, P., Jakubowski, D., & Polakowska, D. (2022). Fluency parameters in the Polish Interpreting Corpus (PINC). In M. Kajzer-Wietrzny, S. Bernardini, A. Ferraresi, & I. Ivaska (Eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament* (pp. 63–91). Language Science Press. https://doi.org/10.5281/zenodo.6977042

Collard, C., & Defrancq, B. (2017, January 12–13). *Sex differences in simultaneous interpreting: A corpus-based study* [Poster session]. The CIUTI Forum 2017, Geneva, Switzerland.

Defrancq, B, Plevoets, K., & Magnifico, C. (2015). Connective items in interpreting and translation: Where do they come from? In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2015* (pp. 192–222). Springer. https://doi.org/10.1007/978-3-319-17948-3

Desmet, B., Vandierendonck, M., & Defrancq, B. (2018). Simultaneous interpretation of numbers and the impact of technological support. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 13–28). Language Science Press. https://doi.org/10.5281/zenodo.1493281

Frittella, F. (2019). '70.6 billion world citizens': Investigating the difficulty of interpreting numbers. *The International Journal of Translation and Interpreting Research*, 11(1), 79–99. https://doi.org/10.12807/ti.111201.2019.a05

Gile, D. (2009). Basic concepts and models for interpreter and translator training (Rev. ed). John Benjamins Pub. Co.

Kajzer-Wietrzny, M. (2012). Interpreting universals and interpreting style. [Ph.D. thesis, Adam Mickiewicz University].

Kajzer-Wietrzny, M., Ivaska, I., & Ferraresi, A. (2021). 'Lost' in interpreting and 'found' in translation: Using an intermodal, multidirectional parallel corpus to investigate the rendition of numbers. *Perspectives*, *29*(4), 469–488. https://doi.org/10.1080/0907676X.2020.1860097

Korpal, P., & Stachowiak-Szymczak, K. (2020). Combined problem triggers in simultaneous interpreting: Exploring the effect of delivery rate on processing and rendering numbers. *Perspectives* 28(1), 126–143. https://doi.org/10.1080/0907676X.2019.1628285

Mazza, C. (2001). Numbers in simultaneous interpretation. *The Interpreters' Newsletter*, 11, 87–104.

Monti, C., Bendazzoli, C., Sandrelli, A., & Russo, M. (2009). Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta*, *50*(4). https://doi.org/10.7202/019850ar

Obler, L. K. (2012). Conference interpreting as extreme language use. *International Journal of Bilingualism*, *16*(2), 177–182. https://doi.org/10.1177/1367006911403199

Pinochi, D. (2009). Simultaneous interpretation of numbers: Comparing German and English to Italian. *The Interpreters' Newsletter*, 14, 33–57.

Seeber, K. (2021, April 28). All just a load of...? A journey into the interpreter's mind. *AIIC*. https://aiic.org/site/blog/the-interpreters-mind

Setton, R. (2011). Corpus-based interpreting studies (CIS): Overview and prospects. In A. Kruger, K. Wallmach, & J. Munday (Eds.), *Corpus-based translation studies: Research and applications* (pp. 33–75). Continuum.

Shlesinger, M. (2002). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta*, *43*(4), 486–493. https://doi.org/10.7202/004136ar

A Softer Image to the Outside World? - Shaping Different Perceptions of China's Image through Interpreting

Binyu *Yang*

The Chinese University of Hong Kong, Shenzhen, China

KEYWORDS: Political Interpreting, CDA, Appraisal Theory, Ideology, Taiwan

Introduction

Nancy Pelosi's visit to Taiwan on August 2, 2022 as the Speaker of the U.S. House of Representatives sparked strong condemnation from Beijing and triggered weeks-long military exercises over the Taiwan Straits, causing disruptions to commercial flights. During the intense weeks leading up to and throughout her visit, the Chinese Ministry of Foreign Affairs (MFA) held regular press conferences with journalists from home and abroad. These conferences provide a significant platform for foreign media to ask questions directly to Chinese authorities, with on-site simultaneous interpreting services provided (Chinese-English bidirectional) for foreign journalists stationed in China.

According to the interpretivist perspective, the world we live in is constructed. Therefore, truth and facts are not fixed or absolute, but rather are often subject to a dynamic and ongoing process of discursive justification (Gu, 2020, p. 40). This is especially true for political discourses. In translational discourses, we cannot assume "that information can circulate unaltered across different linguistic communities and cultures" (Bielsa, 2009, p. 14), because the participation of translators/interpreters will alter the dynamics of the equilibrium of communication. However, how translational political discourses project different national image has been under-explored in the Chinese-English context.

The case study selects the interpreted discourses at the heightened period of political tensions between China and US to investigate different realities (re)constructed by interpreters. It will address the following research questions: (1) What are the major different linguistic features between the ST and TT? (2) How does China construct its image discursively through interpreting? (3) Are there any differences in its portrayal of the national image for domestic and international audiences? If so, how are they different?

Data and Methodologies

To fully account for the narratives before, during, and after Pelosi's visit to Taiwan, the regular press conferences hosted by the Ministry of Foreign Affairs of China covering July 25-August 4 (her visit happened on August 2-3) were selected as the object of the study. The rationale for choosing such a time-frame is that the first questions raised by journalists concerning Pelosi appeared on July 25, and her visit ended on August 3. The data from the chosen time-frame should be sufficient to generate validated results. Only the questions and answers pertaining to Pelosi and her visit were included. All the questions and answers were manually extracted and compiled into Word files before building them into sub-corpora through Sketch Engine (Kilgarriff et al. 2004). The size of the bi-directional corpus is 19,857 tokens, with the Chinese sub-corpus of 8,116 tokens and English sub-corpus of 11,741 tokens. All questions and answers have corresponding translations. Some questions raised by foreign journalists are in English, but all the answers by the Chinese spokespersons are in Chinese.

In terms of the method for the present study, the corpus-based Critical Discourse Analysis (CDA) (Fairclough, 1995) approach is operationalized through two sub-systems of Appraisal Theory- Attitude and Graduation (Martin and White, 2005). These are also the theoretical frameworks for the semantic annotation of corpus data. Critical Discourse Analysis, whose explicit aim is to "make the ideological loading of particular ways of using language and the relations of power that underlie them more visible" (Wodak and Fairclough, 1997, p. 258), provides a better understanding of the translation of ideology by elucidating how discourse shapes and is shaped by ideology (Hatim and Mason, 1997, p. 119). As Baker et al. (2008) rightly highlighted, the combination of discourse analysis and corpus linguistics is a useful methodological synergy.

A quantitative analysis was first conducted. Using Sketch Engine, word frequencies of the both sub-corpora were identified; evaluative words from both sub-corpora were tagged and comparisons were made between ST and TT to identify any salient deviations. For further qualitative analysis, all the evaluative words were again checked manually for accuracy. Major deviations were explained through Appraisal Theory (Martin and White, 2005) and framing strategies (Baker, 2006). Therefore, the whole study is designed in the triangulatory approach, which can function as a means of producing a better (more complete, rigorous, interesting, and/or deeper) analysis (Baker and Levon, 2015).

Preliminary Results

Among the several top frequency words (China, Pelosi, Speaker, Taiwan, Sovereignty), the name *Pelosi* and her title *Speaker* were selected for further analysis because there was a major deviation in the way how Pelosi was addressed by the MFA Spokesperson in ST and TT, which suggests a potential shift in ideology. While the number of *Pelosi* were mentioned equivalent times (both f=96) in both ST and TT, her title, (House) Speaker, was mentioned 39 times in English sub-corpus and only 28 times in Chinese sub-corpus. In other words, Pelosi was addressed more politely in English than in Chinese by the spokesperson. According to sociolinguistic theories, using honorifics title to address the other party not only directly conveys deference and respect, but also signifies relative social status differences (Brown and Gilman, 1968). The interpreted rendition shows more respect to Pelosi and less hostile towards the U.S congress delegation.

In addition, the frequent neutralization and toning down of strong evaluative words regarding China's national image in the interpreted discourse portrays China as less aggressive and hostile to the English-speaking audience as compared to the Chinese audience. Several examples were selected to demonstrate how evaluative items were shifted between ST and TT. The main patterns of these shifts are omission of negative evaluative resources, upscaling of positive attitudinal resources, and selective appropriation. The result of the study shows that discursive (re)construction through interpreting portrays different national images for people within and outside China.

References

Baker, M. (2006). Translation and Conflict: A Narrative Account. Routledge.

Baker, P. K., & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication, 9*(2), 221–236. https://doi.org/10.1177/1750481314568542

Bielsa, E. (2009). Globalization, Political Violence and Translation: An Introduction. In *Palgrave Macmillan UK eBooks (pp. 1–21)*. https://doi.org/10.1057/9780230235410_1

Brown, R. L., & Gilman, A. (1968). The Pronouns of Power and Solidarity. In *De Gruyter eBooks (pp. 252–275)*. https://doi.org/10.1515/9783110805376.252

Fairclough, N. (1995). *Discourse and Social Change*. Polity Press.

Gu, C. (2020). The main problems in China–Japan relations lie in the FACT that some leaders in Japan keep on visiting the Yasukuni Shrine. In B. Wang and J. Munday (Eds), *Advances in discourse analysis of translation and interpreting: Linking linguistic approaches with socio-cultural interpretation (pp. 40–63).* Routledge.

Hatim, B. & Mason, I. (1997) *The Translator as Communicator*. Routledge.

Kilgarriff, A., Rychly, P., Smrz, P. &Tugwell, D. (2004). The Sketch Engine. *Proceedings of Euralex.* Lorient, France, 105-116.

Martin, J. P., & White, P. S. (2005). The Language of Evaluation: Appraisal in English. In *Palgrave Macmillan eBooks.*

Wodak, R. & Fairclough, N. (1997). Critical discourse analysis. In T. A. Van Dijk (ed.), *Discourse as Social Interaction (pp. 258–284)*. Sage

Discourse connectives in German and English translations: An analysis based on automatic annotation and alignment

Frances Yung[1], Merel Scholman[1,2], Ekaterina Lapshinova-Koltunski[3], Christina Pollkläsener[3], Vera Demberg[1]

[1]Saarland University, Germany, [2]Utrecht University, the Netherlands, [3]University of Hildesheim, Germany

KEYWORDS: discourse relations, discourse connectives, German, NLP, explicitation

1. Introduction

Coherence relations that link segments of texts in a discourse can be marked by discourse connectives,such as "because" and "however", or they can be conveyed implicitly. Previous studies show that the translation of discourse connectives depends on various factors. For example, the Explicitation Hypothesis suggests that discourse relations are always explicitated in translation (Blum-Kulka, 1986). Even considering the contrast in explicitness of the source and target languages (with some languagesbeing more prone to expressing discourse relations through explicit connectives than others), researchsuggests that translators tend to insert connectives (Becher, 2011a). Corpus-based studies have identified various factors affecting the explicitation of connectives, such as the constraints and communicative norms of the source and target language, the type of the coherence relations, the connectives involved,and the cognitive interpretability and expectedness of the relations in context (Becher, 2011b; Lapshinova-Koltunski et al., 2022; Marco, 2018; Zufferey & Cartoni, 2014; Zufferey, 2016).

Discourse connectives can be ambiguous on multiple levels; for example, some connectivescan signal multiple relation senses and some connectives can be appear as a connective or with a non-connective function. This can make their identification and classification difficult. Consequently,large-scale corpus studies mostly focus on a handful of connectives and senses. For example , Zufferey and Cartoni (2014) analyzed 200 occurrences each of the English causal connectives *since*, *because* and *given that* in the Europarl Corpus. The frequent causal connective *as* was excluded because it is often used with a non-connective function.

A more comprehensive analysis that takes into account a larger range of connectives and coherence relation senses in the same text is critical to be able to get more insight into the general translation patterns of connectives. To address this, we explore an automatic approach to connective identification and alignment between two different languages. We use automatic discourse parsers to label Englishand German discourse connectives in parallel texts from the Europarl corpus. Furthermore, we use aneural word alignment model to align the words in the texts, to study how the connectives are translated.

Our analysis focuses not only on translation patterns, but also on the feasibility of such an automatic approach. Our research questions include: (1) Are automatic annotations useful for the contrastive study of discourse connectives? (2) What are the translation equivalents of different connectives? (3) How often are different types of connectives explicitated or implicitated?

2. Data and Methodology

The included parallel texts were taken from the Europarl Direct Corpus (Cartoni & Meyer, 2012), which are proceedings from the European Parliament. A total of 33 proceedings are used in the analyses. The data contains 171k tokens of English texts and their German translation from 18 proceedings, and 95K tokens of German texts and their English translation from 15 proceedings[1].

We use two language-specific parsers to annotate the discourse relations in the English and German texts. Discourse connectives in the English original and translated texts are identified by the Discopyparser (Knaebel, 2021), which considers the semantic representation of a connective token and its contexts. The

[1] These 33 and 15 proceedings are selected because they overlap with the discourse-annotated DiscoGem corpus (Scholman et al., 2022), which could be used in future contrastive studies.

classifier distinguishes discourse and non-discourse usage of the connective and labelseach with one of 14 senses, based on the PDTB 2.0 framework (Prasad et al., 2008). The reported accuracies are 97.18% and 86.26% respectively on the tasks of discourse-usage identification and discourse sense classification.

Discourse connectives in the German original and translated texts are similarly identified and classified, using the German Shallow Discourse Parser (Bourgonje & Stede, 2018; Bourgonje, 2021). It has been trained on the Potsdam Commentary Corpus 2.2 (Bourgonje & Stede, 2020) to predict a2nd-level sense labels defined in the PDTB 3.0 hierarchy (16 sense labels[2]) (Webber et al., 2019). Thereported results on the accuracy of this German parser regarding discourse-usage identification and classification are 87.57% and 80.57% respectively.

To study the translation of the discourse connectives, we use the Awesome Align word alignmentmodel (Dou & Neubig, 2021), which extracts corresponding tokens in a pair of bilingual sentencesbased on multilingual embeddings of the tokens. An error rate of 15.1% is reported evaluating againsthuman annotation of English-German word alignments. We manually analyzed 400 random samplesof aligned connective pairs in both translation directions, to ensure that the automatic tools produce reliable annotations on our data.

3. Results

In total, 8058 English and 9739 German connectives have been identified and annotated by the discourse parsers and were aligned automatically. According to our manual analysis, the accuracies of connective identification and 4-way (*EXPANSION*, *CONTINGENCY*, *COMPARISON*, *TEMPORAL*) sense classification are 83% and 91% respectively, and the alignment accuracy is 90%.

We evaluate the mapping of connectives from the source to target texts, as well as from the target to source texts. Among the connectives in the English-original text, 24% of them are not aligned by theword alignment model or are aligned to non-connective words. When looking at the connectives in the German translated text, we find that 40% are not aligned to any connectives in the English source texts, meaning they were *explicitated* or the parser or word aligner was inaccurate. This indicates that whentranslating from English to German, more discourse relations are explicitated than implicitated.

The opposite is observed in the German to English translation direction: 33% of the connectives in the original German texts have been omitted in the English translation, while only 31% of the connectives in the translation have been added.

Evidence of translation-inherent explicitation should take into account the amount of implicitationin the opposite translation direction, since some insertions of connectives are obligatory due to language-specific constraints (Klaudy, 2009, Becher, 2010). From our initial results, we observe that, for both translation directions, the proportions of explicitation are larger than the proportions ofimplicitation in the reverse translation direction.

In addition to explicitation and implicitation through insertion and omission of connectives, we analyzed the cross-lingual alignments of connectives to identify cases where the discourse relations are (under-)specified in the translation. We found similar evidence that discourse relations tend to be more frequently specified in translation than they are under-specified in the opposite translation direction, thus supporting the *Explicitation Hypothesis*.

References

Becher, V. (2010). Towards a more rigorous treatment of the explicitation hypothesis in translation studies. *Trans-kom, 3*(1), 1–25.

[2] Including: *CONJUNCTION*, *INSTANTIATION*, *RESTATEMENT*, *ALTERNATIVE*, *REASON*, *RESULT*, *CHOSEN ALTERNATIVE*, *CONDITION*, *CONTRAST*, *CONCESSION*, *EXCEPTION*, *SYNCHRONY AND SUCCESSION*.

Becher, V. (2011a). When and why do translators add connectives?: A corpus-based study. *Target. International Journal of Translation Studies, 23*(1), 26–47.

Becher, V. (2011b). Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts. PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.

Blum-Kulka, S. (1986). Shifts of Cohesion and Coherence in Translation. In House, J. and Blum-Kulka, S., editors, *Interlingual and intercultural communication*, 17–35. Gunter Narr, Tubingen.

Bourgonje, P. (2021). *Shallow discourse parsing for German* (Vol. 351). IOS Press.

Bourgonje, P., & Stede, M. (2018). Identifying explicit discourse connectives in German. In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, 327–331.

Bourgonje, P., & Stede, M. (2020). The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In Proceedings of the 12th Language Resources and Evaluation Conference, 1061–1066.

Cartoni, B., & Meyer, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2132–2137.

Dou, Z.-Y., & Neubig, G. (2021). In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2112–2128.

Klaudy, K. (2009). The asymmetry hypothesis in translation research. Translators and their readers. In Homage to Eugene A. Nida, 283–303. Brussels: Les Editions du Hazard.

Knaebel, R. (2021). discopy: A neural system for shallow discourse parsing. In Proceedings of the 2nd Workshop on Computational Approaches to Discourse, 128–133.

Lapshinova-Koltunski, E., Pollkläsener, C., & Przybyl, H. (2022). Exploring explicitation and implicitation in parallel interpreting and translation corpora. The Prague Bulletin of Mathematical Linguistics, (119), 5–22.

Marco, J. (2018). Connectives as indicators of explicitation in literary translation: A study based on a comparable and parallel corpus. Target. International Journal of Translation Studies, 30(1), 87–111.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008). The Penn Discourse treebank 2.0. In LREC 2008.

Scholman, M., Tianai, D., Yung, F., & Demberg, V. (2022). Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022), 3281–3290. European Language Resources Association.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). The Penn discourse treebank 3.0 annotation manual, (35), 108, Philadelphia: University of Pennsylvania.

Zufferey, S. (2016). Discourse connectives across languages: Factors influencing their explicit or implicit translation. *Languages in Contrast: International Journal for Contrastive Linguistics, 16*(2), 264–279.

Zufferey, S., & Cartoni, B. (2014). A multifactorial analysis of explicitation in translation. *Target: International Journal of Translation Studies, 26*(3), 361–384.

Acquis terminology in English-Georgian translation

Khatuna Beridze

Batumi Shota Rustaveli State University, Georgia

KEYWORDS: Acquis Communautaire, Parallel Corpus, EU-Georgia

Translation of the Acquis Communautaire was followed in almost every acceding EU state with debates about the quality, multiple cases of requests for corrigenda, and controversies regarding the terminology consistency. Insignificant in one respect, or undetected in another, translational mistakes incurred costs as well as additional efforts on behalf of the EU and national institutions. The paper examines consistency of terminology of the EU-Georgian Association Agreement, Title IV, articles 1-276. The analyses were conducted using a custom-built bilingual corpus platform and software. The goal of this research was to continue development of the CT analyses and methodology for the translated Acquis, as well as legal texts in general. Based on the analyzed samples, the study categorizes 8 linguistic approaches to terminology translation in the legal texts that cause inconsistency and issuant teleological misinterpretation of the law. Another value of the research is that it ascertains and promotes the significance of the corpuslinguistic methodology for the development of the CTS of the legal texts. In the first stage of the research, we manually tagged collocations, calques, foreign words, terms (legal and economic), syntactic calques, misinterpretation of the passive constructions, translation errors, abbreviations, lexical additions. The manual processing of the texts was carried out at three stages, each stage demands: 1. pre-processing comparative analysis of the parallel AA texts; 2. manual alignment; 3. accurate performance of the tagging process; 4. supervision at each stage and before the input is uploaded into the experimental parallel corpus analyzer. The body is encrypted in XML format, in accordance with the Basic Text Encryption Recommendations (TEI P4). Texts are in one language form a single TEI body, consisting of the text itself and the TEI title (a header). Each text in turn contains a TEI title and its own text. The corpus directionality is unidirectional – English to Georgian. The software automatically aligns bilingual AA texts, and extracts aligned bilingual EU-Georgia Association Agreement Terminology Lexicon. The software a. analyzes the texts in XML format; b. outputs statistics, compares and visually displays the data. c. extracts the ST and TT legal term pairs. The qualitative analysis relies on the quantitative data obtained from the corpus. The methodology used for this research is based on the semi-automatic linguo-statistical data eliciting from the bilingual corpus, using pre-determined TagSet, custom-built corpus platform and software. In accordance with the objectives of the research, we examined the quantitative data to assess the quality of translation, and categorized translation inconsistencies of the AA Title IV, articles 1-276. Part of the study is presented in the paper, which outlines categories of terminology mistranslations.

Putting discourse in political context: a case study of Taiwanese identity

Yu-Shen Cheng

Univesitat Pompeu Fabra, Spain

KEYWORDS: corpus-based discourse analysis, transitivity, presidential speeches, Taiwanese identity, mixed-effects linear model

The evolution of the construction of Taiwanese national identity can be viewed as the reflection of the transition from an authoritarian dictatorship (1949-1991, when Temporary Provisions against the Communist Rebellion were officially ended) dominated by Kuomintang, or Chinese Nationalist Party, to a more locally focused democratically elected government. During the dictatorship, the Chinese identity was reinforced in the education in order to propagate the anti-communist ideology and enhance the status of Han culture over others, while after the plea of bottom-up movements starting from mid-1990s, the society witnessed a significant growth of Taiwanese Consciousness, which is encouraged by the succession of the first non-KMT president in 2000 (Hung, 2016).

According to De Cillia, Reisigl et al. (1999), discourses can be used to construct national identity for their possible abilities to start, produce and fabricate certain social conditions in which politicians, intellectuals, and media reify the notion of a specific national community.

Presidential addresses, as a highly ideologically and politically oriented discourse, may reveal relevant ideological positions realized covertly through the selection of linguistic choices (Kpeglo and Giddi, 2022), as well as methods adapted by presidents to clarify their stand toward major political issues and establish closer relationships to their supporters (Kondowe, 2014; Chen, 2018; Cao and Lei, 2022). In the specific political situation of Taiwan, the presidents, apart from internal and external issues, would also tackle the Chinese/Taiwanese identity in their addresses.

This research sets out to investigate how the changing political conditions where National Day speeches of the 8th to 15th Taiwanese presidents, in other words who come into power through direct elections, are addressed affect the process types and participant roles assigned to the citizens based on transitivity analysis.The transitivity system forms part of the lexicogrammatical realization of ideational function at the clause level, where the experience is codified by process, participants, and associated circumstances. As Hart (2014: 22) explains further, the process, as a core constituent in the clause, represents the choice determined in the discourse and involves the identification of participant roles in a certain event.

Regarding national identity and transitivity system, Alamenda-Hernández (2007) and Alaghbary (2017) conclude that process types and participant roles designated may differ in the discourse, which in turn designate positive or negative images of a certain community. Based on this observation, this research analyses how the citizenship is presented through process types and participant roles in inaugural addresses.

This investigation hypothesizes that participant roles of the people which are determined by process types show different tendencies between discursive references as Taiwanese and Chinese, which reflect social and ideological changes regarding to the identity issue. In order to examine this hypothesis, this study utilizes two different data, a manually annotated corpus according to the transitivity system and self-reported surveys retrieved from Trends in Core Political Attitudes among Taiwanese (Election Study Center of National ChengChi University, 2022).

Corpus-based approach is a useful method to discover the patterns of how participant roles are assigned in comparable discourse for further interpretation about the results (Lee, 2016). Therefore, corpus is applied in this research with an intention to operationalize the results in statical model to study the relation between discourse and context. The transitivity system also permits the receptors to identify the implicit relations between the participants in the clauses while the texts describe the same situation and event (Ammara, Anjum & Maryriam, 2019). The corpus built specifically for this investigation is composed of

the transcriptions in Mandarin Chinese of the 8th to the 15th Taiwanese presidents' National Day addresses to examine the process types and the participant roles used to refer to the citizens.

For example, in the following phrase extracted from the National Day address in 1997, "We can proudly say that Republic of China has realized Mr. Sun Yat-sen's ideals in Taiwan, Penghu, Kinmen and Matsu […] and writes a new page in the history for the Chinese nation", the first plural pronoun and the Chinese nation are used to refer to the same target, the Taiwanese citizens, while being assigned to different roles in the discourse, respectively as sayer and the goal according to transitivity system. In the same year, 41.4% of the people reported themselves as both Taiwanese and Chinese and 34% as solely Taiwanese according to the project conducted by NCCU.

The context where the discourse is produced is operationalized as the following topics, Chinese/Taiwanese Identity and Independence vs. Unification with the Mainland, which forms part of the annual survey conducted by National ChengChi University. Process types and participant roles will be predicted by Chinese/Taiwanese Identity and Independence vs. Unification with the Mainland, as fixed effects while the year of address and the interaction between the two predictors will be included as random effects in mixed-effects linear regression models for its ability to provide more reliable estimation about the distribution of the dependent variables (Coupe, 2018).

Corpus-based approach to transitivity analysis allows large-scale observation of linguistic patterns of process types implemented between different presidents and participant roles assigned to the 'Taiwanese' and similar notions referring to the citizenship at the grammatical level. On the other hand, the political context where the discourse is produced is also an analytical focus in this research, which is measured by individual attitudes about national identity and relevant issues, in order to provide an explanation of the linguistic patterns at the pragmatical level.

This research proposes an alternative method to analyze the discourse in the context where it is produced. The combination of corpus-based approach and survey data collection enables a new analytical perspective on the results from corpus-based discourse analysis, adopting a scientific procedure to interpret how contexts can possibly affect the discursive patterns while the contexts are defined and measured statistically.

References

Alameda-Hernández, A. (2008). SFL and CDA: Contributions of the Analysis of Transitivity System in the Study of the Discursive Construction of National Identity (Case study: Gibraltar), *Linguistics Journal*, 3:3, pp.169-175, Linguistics Journal Press.

Alaghbary, G. S. (2017). Identity and National Belonging in Ansaruallah's Political Rhetoric: A Transitivity Analysis, *International Journal of English Linguistics*, 7:4, Canadian Center of Science and Education.

Ammara, U., Anjum, R. Y. & Javed, M. (2019). A Corpus-based Halliday's Transitivity Analysis of 'To the Lighthouse', *Linguistics and Literature Review*, 5(2), pp. 139-162.

Cao, X. and Lei, X. (2022). Critical Discourse Analysis of Inaugural Speeches by Trump and Biden Based on the Systemic Functional Grammar, *International Journal of Liberal Arts and Social Science*, 10 (1), pp.1-16.

Chen, W. (2018). A Critical Discourse Analysis of Donald Trump's Inaugural Speech from the Perspective of Systemic Functional Grammar, *Theory and Practice in Language Studies*, 8(8), pp.966-972.

Coupé, C. (2018). Modeling Linguistic Variables With Regression Models: Addressing Non-Gaussian Distributions, Non-independent Observations, and Non-linear Predictors With Random Effects and Generalized Additive Models for Location, Scale, and Shape, *Frontiers in Psychology*, 9, pp.513-513.

De Cillia, R., Reisigl, M. and Wodak, R. (1999). The discursive construction of national identities, *Discourse & Society*, 10:2, pp.179-173. Sage.

Election Study Center, National Chenchi University (20022). Trends in Core Political Attitudes among Taiwanese. Retrieved in 2022 from https://esc.nccu.edu.tw/PageDoc/Detail?fid=7800&id=6961.

Halliday, M.A.K. & Matthiessen C.M.I.M. (1999). Construing Experience through Meaning: A Language-based Approach to Cognition. London: Cassell.

Hart, C. (2014). Discourse, Grammar and Ideology – Functional and Cognitive Perspectives. Bloomsbury.

Hung, C.Y. (2016). Ambiguity as deliberate strategy: the 'de-politicized' discourse of national identity in Taiwanese citizenship curriculum, *Critical Studies in Education*, 57:3, pp.394-410. Routledge.

Kondowe, W. (2013). Interpersonal Metafunctions in Bingu wa Mutharika's Second-Term Political Discourse: *A Systemic Functional Grammatical Approach*, *International Journal of Linguistics*, 6 (3), pp.70-84.

Kpeglo, S.B, and Giddi, E. (2022). Ideological Positioning in Presidential Inaugural Addresses: A Comparative Critical Discourse Analysis of President Agyekum Kufour and President John Evans Ata Mills, *Integrated Journal for Research in Arts and Humanities*, 2 (5), pp.92-103.

Lee, C.S. (2016). A corpus-based approach to transitivity analysis at grammatical and conceptual levels: A case study of South Korean newspaper discourse, *International Journal of Corpus Linguistics*, 21(4), pp.465-498. John Benjamins Publishing Company.

Individual pause measures as possible indicators of unchallenged translation in the TPR-DB

Michiel Kusé, Evy Woumans

Ghent University, Belgium

KEYWORDS: Cognitive Translation Studies, Translation Process Research, Translation Process Research Database, keylogging

Throughout our daily lives, we rely on a sort of automatic pilot for a great number of activities; easy examples tend to include a motor component, but language processing is often no less like riding a bike. The apparent ease with which we perform cognitively demanding activities such as reading and writing, including in second and third languages, has raised the question if such smooth processing can also be found in the translation process, which has in turn given rise to the 'default translation' construct. Default translation can be defined as "a particular phase of translation production [in which] translators demonstrate stretches of uninterrupted production" (Halverson, 2019, p. 190) and in which processing occurs relatively automatically and with limited cognitive effort (Carl & Dragsted, 2012).

This study is part of a larger project that aims to investigate what textual features spur the occurrence of default translation patterns in both the translation and interpreting process, and how this occurrence is influenced by training and experience on the one hand, and by linguistic distance on the other. In this first study, we applied the methodology suggested by Halverson (2019) to parts of the CRITT Translation Process Research Database (TPR-DB). Given the absence of gaze data in most of this corpus, we did not go so far as to identify default translation patterns, but did identify stretches of 'unchallenged' target text (TT) production (Carl & Dragsted, 2012).

Our corpus consisted of keylogging data gathered during from-scratch translation projects, with a set of six English texts as the source, and Dutch (10 participants, 60 texts; Vanroy, 2021), German (24 participants, 47 texts; Carl, Jakobsen, & Schaeffer, 2015), and Spanish (32 participants, 60 texts; Mesa-Lao, 2012) as target languages. For each of the subjects within our corpus, individual pause thresholds were determined based on recommendations by Rosenqvist (2015), resulting in a lower and upper threshold. This resulted in four pause types: 'delay' (under 200 ms), 'short' (200 ms – lower threshold), 'medium' (lower - upper threshold), and 'long' (over the upper threshold), with production units (PUs) defined as any stretch of TT production between two long pauses. Following Muñoz Martín and Cardona Guerra (2018), we identified 'broken words', i.e. short or medium pauses within a word, or 'changes', i.e. modifications to the existing target text, allowing us to identify problems and thereby challenged, unchallenged, and empty production units.

Preliminary findings indicate that unchallenged PUs occur evenly across all texts, taking up around 25 % of keystrokes, but show large inter- and intra-individual variation. Unchallenged PUs also proved significantly shorter than challenged or even than the average PU. Furthermore, they are more likely to occur around sentence boundaries, as are longer pauses – possibly these function as natural resting points and allow planning of the upcoming TT production, facilitating it. Additionally, unchallenged production was significantly less likely when translating into Spanish, and significantly more so when translating into German. However, specific source text characteristics did not have any effect on the occurrence of unchallenged target text production.

References

Carl, M., & Dragsted, B. (2012). Inside the monitor model: Processes of default and challenged translation production. *Translation: Corpora, Computation, Cognition, 2*(1), 127–145.

Carl, M., Jakobsen, A.L., & Schaeffer, M. (2015, January 29–30). *Literality in Translation and Processes of Post-Editing* [Conference presentation]. Translation in Transition, Germersheim, Gemany.

Halverson, S. L. (2019). 'Default' translation: A construct for cognitive translation and interpreting studies. *Translation, Cognition & Behavior, 2*(2), 187–210.

Mesa-Lao, B. (2012, August 17-18). *Translating vs Post-editing: A pilot study on eye movement behaviour across source texts* [Conference presentation]. International Workshop on Expertise in Translation and Post-editing - Research and Application, Copenhagen Business School, Denmark.

Muñoz Martín, R., & Cardona Guerra, J.M. (2018). Translating in Fits and Starts. Pause thresholds and roles in the research of translation processes. *Perspectives: Studies in Translation Theory and Practice, 27*(4), 525–551.

Rosenqvist, S. (2015). *Developing Pause Thresholds for Keystroke Logging Analysis* [Bachelor paper]. University of Umeå, Sweden. Available at https://www.diva-portal.org/smash/get/diva2:834468/FULLTEXT01.pdf

Vanroy, B. (2021). *Syntactic difficulties in translation* [Doctoral dissertation, Ghent University]. Ghent University Biblio. https://biblio.ugent.be/publication/8709063/file/8718358.pdf

Comparison of Semantic Features between Translated and Original Chinese Literary Texts: from a Diachronic Perspective

Pang Shuangzi[1], Wang Kefei[2]

[1]Shanghai Jiao Tong University, China, [2]Beijing Foreign Studies University, China

KEYWORDS: semantic features, translated Chinese, original Chinese, literary, diachronic

For a long time, Corpus-based Translation Studies has primarily focused on easily identifiable and marked linguistic structures, with limited exploration of semantic levels due to the challenge of empirically operationalizing indicators. However, recent studies have aimed to investigate the semantic features in translated texts. For example, Vandevoorde (2017) employed the 'semantic mirror' method to examine the semantic elements of the verb "start" in translated and non-translated texts, revealing that translated texts tend to exhibit flattened semantic meaning compared to non-translated texts. Similarly, De Baets Pauline et al. (2018) used the 'behavioral profile' approach to investigate semantic features in translated texts and verify the 'semantic stability' hypothesis. In China, several studies on semantic features have also highlighted the importance of corpus-based research in Translation Studies, but most of these studies have been limited to synchronic analysis and random case selection.

To address this research gap, this paper aims to dynamically investigate the semantic features of translated Chinese literary texts in comparison to the original Chinese texts. The study intends to answer the following questions using multivariate statistics: (1) Are the semantic features demonstrated in translated Chinese literary texts different from the original Chinese literary texts from a diachronic perspective? (2) Which semantic categories most distinguish the three sampling periods in translated Chinese texts? (3) To what extent do semantic features change over time compared to lexical and grammatical features?

To conduct this study, we utilize the Synonym Dictionary (1996) written by Mei Jiaju to semantically annotate the Chinese Diachronic Composite Corpora (CDCC). These corpora, designed to explore the relationship between language contact and translation, consist of a parallel corpus and comparable corpora collected during three periods: 1927–1937, 1956–1962, and 1987–1997. The corpus comprises various literary text types over 9 million printed tokens, with a diachronic parallel corpus of 5,791,871 tokens and a diachronic comparable corpus of 6,688,525 tokens(Pang and Wang, 2020). The corpus has been annotated using The University of Pennsylvania Treebank Tag-set (Santorini 1990) for English texts and ICTCLAS in the Chinese Academy of Sciences (Zhang & Liu 2002) for Chinese texts. The parallel corpus has been aligned at the sentence level.

In this study, we employ the *Synonym Dictionary* (1996) written by Mei Jiaju to semantically annotate the CDCC (Chinese Diachronic Composite Corpora), which are designed to examine the relationship between language contact and translation, including a parallel corpus and comparable corpora collected in three periods: 1927–1937, 1956–1962 and 1987–1997. The corpus comprises just over 9 million printed tokens, with a diachronic parallel corpus of 5,791,871 tokens and a diachronic comparable corpus of 6,688,525 tokens. Various literary text types are represented across the corpora, organized into nine genres: religious writings, biography, essays, fairytales, general fiction, romantic fiction, adventure stories, science fiction and humorous fiction(Pang & Wang, 2020).The corpus has been annotated using The University of Pennsylvania Treebank Tag-set (Santorini 1990) for the English texts and ICTCLAS in Chinese Academy of Sciences (Zhang & Liu 2002) for the Chinese texts, and the parallel corpus has been aligned at sentence level.

The Synonym Dictionary (1996) operates on a framework comprising 12 semantic categories: A. human beings, B. objects, C. time and space, D. abstract matters, E. features, F. actions, G. mental activities, H. activities, I. phenomena and states, J. relationship, K. auxiliaries, and L. the complimentary close. Originally developed at Harbin Industry University in China, this system has predominantly been employed in the domain of national language processing in previous studies. However, this paper seeks to

expand the system's application to translation studies, specifically investigating semantic differences in translation from a diachronic perspective.

The study adopts a three-step methodology. Firstly, an analysis of diachronic parallel corpora is conducted to examine the distinctive features present in translated Chinese literary texts. Secondly, by analyzing diachronic comparable corpora, a comparison is made between Chinese translated Chinese literary texts from three sampling periods and their original Chinese counterparts, aiming to identify potential disparities in semantic features between them. Finally, in order to assess the extent of change in semantic features relative to lexical and grammatical items, a collection of 42 lexical-grammatical features in Chinese, based on Biber's (1988) model, is utilized. The overall degree of change in these features is evaluated using the random forest method.

Based on the statistical results, it is observed that translated Chinese texts exhibit distinct features in semantic types K (auxiliaries), H (activities), E (features), and D (abstract matters), as these types rank highest in frequency among all the semantic categories. Specifically, the frequencies of semantic types in translated Chinese literary texts can be ordered as follows: auxiliaries > activities > features > abstract matters > relationship > objects > human beings > phenomena and states > time and space > action > mental activities > the complimentary close.

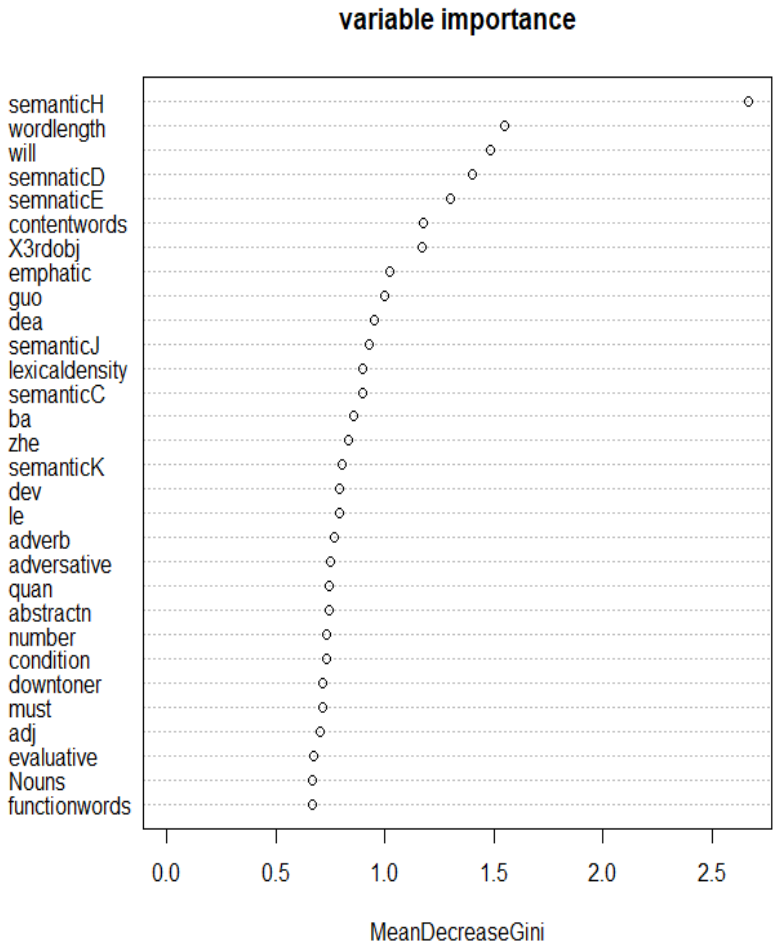


Figure 1. Variable importance plot: The relative importance of indicator scores in predicting the period of CDCC

In conclusion, the analysis of translated Chinese literary texts compared to their original counterparts reveals significant changes in semantic features across different periods in the corpus (refer to Figure 1). The results obtained from the random forest analysis indicate that when all indicators are considered

collectively, semantic types H, D, and E emerge as the most distinct in distinguishing the various periods in the corpus. This finding supports the notion that semantic change holds greater prominence than lexical and grammatical changes throughout the three examined periods, which further extends the assumption proposed by Kranich (2011, p. 13).

Additionally, our study highlights that translated Chinese literary texts exhibit similar changing trends to the original Chinese literary texts in terms of the frequencies observed in human beings, time and space, features, mental activities, relationship, and auxiliaries. However, when considering the overall utilization of the 12 semantic categories, translated Chinese literary texts significantly differ from the original Chinese literary texts. This suggests that, from a diachronic perspective, semantic features tend to remain more stable in translated Chinese literary texts compared to their original counterparts.

References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

De Baets, P., Vandevoorde, L., & De Sutter, G. (2018). Meaning Shifts in Translation: A Corpus-based Behavioural Profile Approach. In S. Granger, M.-A. Lefer, & L. Penha-Marion (Eds.), *Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference* (5th edition) (pp. 47-49). Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain.

Kranich, S., Becher, V., & Höder, S. (2011). A Tentative Typology of Translation-induced Language Change. In S. Kranich, V. Becher, & S. Höder (Eds.), *Multilingual Discourse Production: Diachronic and Synchronic Perspectives* (pp. 9-44). Amsterdam: John Benjamins Publishing Company.

Mei, J. (1996). *Synonym Dictionary*. Shanghai: Shanghai Lexicographical Publishing House.

Pang, S., & Wang, K. (2020). Language Contact through Translation: The Influence of Explicitness in English-Chinese Translation on Language Change in Vernacular Chinese. *Target,* 42(3), 420-455.

Vandevoorde, L., Lefever, E., Plevoets, K., & De Sutter, G. (2017). A corpus-based study of semantic differences in translation: The case of inchoativity in Dutch. *Target,* 29(3), 388-415.

Zhang, H., & Liu, Q. (2002). Chinese Lexical Analysis System (ICTCLA). Institute of Computing Technology. Retrieved March 12, 2020.

Translating the phraseology of Polish academic legal language into English. Insights from corpora of translated and non-translated academic legal writing in English

Anna Setkowicz-Ryszka

University of Łódź, Poland

KEYWORDS: legal scholarship, academic legal language, fixed phrases, Latin maxims

Legal translation is a kind of specialized translation which is recognized as fraught with particular difficulties, including inherent problems with incongruence between concepts from various legal systems (Doczekalska, 2013; Jopek-Bosiacka, 2013; Šarčević, 1997). However, the differences go as deep as the general way of writing, or "unstated conventions (Cao, 2007). Thus legal translators need awareness of and ability to address asymmetry between concepts from different legal systems, as well as knowledge of specialist genre conventions, legal registers, legal collocations, and terminology in both source and target languages (Popiołek, 2020).

Academic legal writing is sometimes distinguished as a separate kind of legal discourse. As globalization has reached academia, legal scholars are required to publish in English, at least titles and abstracts of their articles. Apart from the usual challenges of legal translation, the language used by legal scholars merits attention due to the presence fixed phrases, including Latinisms. Pontrandolfo finds that phraseology is a major obstacle for legal translators and that there are few resources they can consult (2015). The analysed phrases belong to a category which Kjaer calls "routine phrases whose use is merely habitual", which are not required by any norms, but reduce "the immense burden of text production" (2007).

The aim of this corpus-based study is to identify potential equivalents of fixed phrased used by Polish legal scholars – selected on the basis of the author's professional practice and the list of Latin maxims in Gałuskina & Sycz (2013) – in the law and political science sections of:

- Directory of Open Access Journals (DOAJ) corpus
- Oxford Corpus of Academic English (OCAE)
- Corpus of Contemporary American English (COCA)
- British National Corpus (BNC).

The phrases include purely Polish ones, like: *w doktrynie i orzecznictwie*, mixed phrases: *postulaty de lege ferenda, wyrazić expressis verbis,* and Latin maxims: *nulla poena (sine lege)* or *ne bis in idem*. The analysis covers mainly comparing frequencies and reading concordance lines in the consulted corpora and, where possible, frequent collocates of selected phrasemes. DOAJ corpus is assumed to contain mainly translated texts and the remaining corpora are more likely to contain non-translated texts.

The analysis reveals differences in the frequencies of phrases of interest in translated and non-translated texts. Translated texts display higher frequencies of literal translations of the analysed phrases (*doctrine and jurisprudence/case law*). Latin expressions are often left untranslated (*mention expressis verbis*) or not changed to nominative form (*de lege ferenda*), disregarding their actual use in native English (Setkowicz-Ryszka, 2022). But without knowing who reads translated legal scholarship from civil-law countries, it is hard to recommend whether conventions of common law scholarship should be followed or whether features of English as a Lingua Franca, like unusual linguistic patterns or cross-linguistic influence (Albl-Mikasa et al., 2020), should be tolerated as more understandable to speakers of other languages of civil law area.

References

Albl-Mikasa, M., Ehrensberger-Dow, M., Hunziker Heeb, A., Lehr, C., Boos, M., Kobi, M., Jäncke, L., & Elmer, S. (2020). Cognitive load in relation to non-standard language input: Insights from interpreting, translationand neuropsychology. *Translation, Cognition & Behavior*, *3*(2), 263–286. https://doi.org/https://doi.org/10.1075/tcb.00044.alb

Alcaraz Varó, E., & Hughes, B. (2002). *Legal Translation Explained*. St.Jerome Publishing.

Cao, D. (2007). *Translating law*. Multilingual Matters.

Doczekalska, A. (2013). Comparative law and legal translation in the search for functional equivalents - Separate or intertwined domains. *Comparative Legilinguistics*, *16*, 63–75.

Gałuskina, K., & Sycz, J. (2013). Latin maxims and phrases in the Polish, English and French legal systems - the comparative study. *Studies in Logic, Grammar and Rhetoric*, *34*(47), 9–26. https://doi.org/DOI: 10.2478/slgr-2013-0020

Jopek-Bosiacka, A. (2013). Comparative law and equivalence assessment of system-bound terms in EU legal translation. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 110–146.

Kjær, A. L. (2007). Phrasemes in legal texts. In H. Burger, D. Dobrovol'skij, P. Kühn, & N. R. Norrick (Eds.), *Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research* (pp. 506–516). Walter de Gruyter.

Pontrandolfo, G. (2015). Investigating Judicial Phraseology with COSPE: A contrastive Corpus-based Study. In *New Directions in Corpus-Based Translation Studies* (pp. 137–159).

Popiołek, M. (2020). ISO 20771 : 2020 overview and legal translator competence requirements in the context of the European Qualifications Framework , ISO 17100 : 2015 and relevant research. *Lingua Legis*, *28*, 7–40. https://lingualegis.ils.uw.edu.pl/index.php/lingualegis/article/view/60/49

Šarčević, S. (1997). *New Approach to Legal Translation*. Kluwer Law International.

Setkowicz-Ryszka, A. (2022). Wyrażenia łacińskie w polskich i angielskich tekstach prawniczych z perspektywy tłumacza. In E. Kujawska-Lis, A. Krawczyk-Łaskarzewska, & O. Letka-Spychała (Eds.), *Komunikacja międzykulturowa w świetle współczesnej translatologii. Tom IX. Języki, teksty, interpretacje* (pp. 219–232). Wydawnictwo UWM.

A collocation-based analysis across English and Italian. The case of migrant and migration in international, Italian and UK migration law

Virginia Zorzi and Cinzia Spinzi

University of Bergamo, Italy

KEYWORDS: migration law; migration discourses; contrastive analysis; corpus-based analysis; collocation analysis

Discourses on migration represent and reproduce the way societies view and make decisions on this phenomenon. The language employed in institutional settings to regulate migration is especially relevant, since migration is directly affected by it. Migration is regulated by international laws, treaties and conventions at different levels, (e.g. UN, EU), as well as by national laws. Each layer of legislation is tied to a different context and policy framework. In any of these contexts, people who migrate and their experiences need to be described through specific terms, which are not always consistent across different institutions and may not be translated literally. While this may assure an appropriate distance between international and local stances (Peruzzo, 2012), it could also cause ambiguities and misunderstandings (Corbolante, 2015). Nor can these terms – as any instance of language in use – be considered completely neutral and objective; in fact, they have sometimes been claimed to dehumanise (Loupaki, 2018) and criminalise migrants (Pace & Severance, 2016).

Adopting a corpus-based, contrastive perspective in the study of migration law discourse (cf. Sanchez et al., 2019), the present study focuses on how the terms "migrant" and "migration", selected as the most generic terms used to refer to people on the move, are used across languages and contexts. Geographically speaking, it concentrates on the European geographical area, with special attention to Italy and the UK. More precisely, it answers the following research questions:

(1) How are "migrant" and "migration" used in official texts regulating migration at the international level? More specifically, what semantic domains are associated with these terms? Are there differences between UN and EU texts?

(2) How are these terms used in Italian official texts regulating migration and what are those that could be considered equivalent to "migrant" and "migration"? How are these terms used in English-language migration legislation applying to the whole UK? Does UK legislation include any other equivalent terms?

In order to answer the above questions, a corpus-based analysis of official documents regulating migration, published after WWII and currently in force, was carried out. Three corpora were collected:

1) an 861,725-word English-language corpus, comprising international migration law documents (the English-language International Migration Law Corpus). It includes texts issued by the UN and its organisations (International Organization for Migration, UNHCR, International Labour Organization, International Maritime Organization) and by the EU and the Council of Europe. This corpus can be divided into two sub-corpora: the UN Migration Law Corpus (338,354 words) and the EU Migration Law Corpus (523,371 words);
2) a 404,664-word Italian-language corpus, comprising national migration law documents (the Italian Migration Law Corpus).
3) an 888,298-word English-language corpus, comprising Public General Acts and Statutory instruments on migration applying to the whole UK (the UK Migration Law Corpus).

Relevant documents were identified with the help of reference texts (Plender, 2007; Chetail, 2014) and online web pages and databases, and collected manually from said sources. The three corpora are comparable in terms of topic, text type and time span, but differ in scope (international or national).

Drawing on Corpus Linguistics tools and using the software #LancsBox (Brezina et al., 2021), we carried out an analysis of collocates of the two lemmas "MIGRANT" and "MIGRATION" (lemmas henceforth marked in uppercase) in the international corpus. Collocates were extracted with a span of five words to the left and right of node words, and MI as the collocation strength measure (minimum threshold set at 3, cf. Hunston, 2002: 71). Minimum collocation frequency was set at 1% of the frequency of each node word, and never below 5; when node word frequencies were too low, concordances were inspected to identify any relevant patterns. Collocates were examined individually through concordance inspections and aggregated into semantic domains identified by the analysts (cf. Baker et al, 2013), which allowed for a comparison between UN and EU documents. Subsequently, a contrastive analysis was carried out, which uncovered non-equivalence of the two terms between the international and national corpora. Indeed, "MIGRANT" and "MIGRATION" are much less frequent in the UK corpus than in the international corpus, as are their prima facie Italian translations in the Italian corpus. Therefore, Manca's (2004; 2012) method was followed: (1) the collocates of the word(s) under consideration were identified; (2) equivalents to these collocates were searched in the Italian corpus; (3) collocates of these equivalents were analysed, looking for potential equivalents to the word(s) initially under consideration. A similar method was applied to the UK corpus, where the collocates of the collocates extracted in (1) were analysed in search for terms with similar collocational profiles as those of "MIGRANT" and "MIGRATION" in the international corpus.

Preliminary results show that the semantic field appearing most frequently among the collocations of "MIGRANT" in the international corpus is that related to international protection and support for migrants, followed by migrant work/labour. Collocations referring to protection/support and work/labour are the most frequent ones also in the UN sub-corpus, while contrasting smuggling and trafficking seems to be the domain most often associated with migrants in the EU sub-corpus. As for the lemma "MIGRATION", collocations referring to migration management are the most frequent ones in the international corpus, as well as in both its sub-corpora.

In order to identify equivalent(s) of "MIGRANT" in the Italian and UK corpora, its four most frequent collocates in the international corpus ("workers", "rights", "members", "irregular") were used to apply the collocation-based identification method described above. The results indicate that "straniero" (foreigner, foreign, alien) is used as an equivalent of "MIGRANT" in Italian (cf. Taylor, 2009: 15 on UK vs Italian news). In the UK corpus, "citizens" emerged as a potential equivalent to "MIGRANT", although concordance analysis revealed that its use is circumscribed to citizens of certain (mainly EU) countries, while the relatively few occurrences of "MIGRANT" were mostly associated to the UK Tier-based migrant classification system. The same procedure was applied to the four most frequent collocates of "MIGRATION" in the international corpus ("management", "asylum", "international", "regulation"), with "immigrazione" (immigration) in the Italian corpus and "immigration" in the UK corpus emerging as potential equivalents.

Taking into account both intra-linguistic and cross-linguistic differences through the four-fold corpus design (UN, EU, Italy, UK), this research seeks to account for the conceptual complexity of migration-related terminology and discourse even beyond diverging or ambiguous definitions. Our preliminary results confirm this complexity, suggesting that the concepts of migrant and migration as used in UN versus EU documents are partly associated with different domains. Moreover, the analysis seems to highlight some out-grouping tendencies characterising Italian and UK data, where texts appear to predominantly frame migrants from the perspective of in-group members (i.e. Italian or EU citizens) through specific terminological choices. Thus, the mismatch emerging between the English terms used in the international corpus and their literal, technically accurate translations "MIGRANTE" and "MIGRAZIONE" points to different potential translation equivalents used in the actual context of migration legislation in Italy. A similar mismatch appears between the international and UK corpora. While responding to the clear need of distinguishing citizens and non-citizens of a country or of EU countries, this mismatch also uncovers long-standing attitudes underlying the representation of people on the move in this context. These results highlight the importance of discussing how these terms are to be used and translated, not only in institutional settings, but also in non-specialised public debates.

References

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the British press 1998–2009. *Applied linguistics, 34*(3), 255-278.

Brezina, V. (2018). Statistics in corpus linguistics: A practical guide. Cambridge University Press.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 6.x. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox. (23/08/2022)

Chetail, V. (2014). *International migration law*. Oxford University Press.

Corbolante, L. (2015, September). Le differenze tra rifugiati e migranti. From the website "Terminologia etc. Terminologia, localizzazione, traduzione e altre considerazioni linguistiche". http://blog.terminologiaetc.it/2015/09/03/significato-migrante-rifugiato-ue-vs-unhcr/ (30/11/2022).

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Loupaki, E. (2018). EU Legal Language and Translation-Dehumanizing the Refugee Crisis. *International Journal of Language & Law, 7*, 97-116.

Manca, E. (2004). Translation by Collocation. The Language of Tourism in English and *Italian*. TWC

Manca, E. (2012). Translating the Language of Tourism across Cultures: from functionally complete units of meaning to cultural equivalence. *Textus, 25*(1).

Pace, P., Severance, K. (2016). Migration terminology matters. *Forced Migration Review, 51*, 69-70.

Peruzzo, K. (2013). Terminological equivalence and variation in the EU multi-level jurisdiction: a case study on victims of crime. Phd Thesis, University of Trieste.

Plender, R. (ed.). (2007) *Basic Documents on International Migration Law*. Martinus Nijhoff.

Sanchez, P., Aguado, P., Perez-Paredes, P. (2019). Featuring immigrants and citizens A comparison between Spanish and English primary legislation and administration information texts (2007-2011). In Viola, L., & Musolff, A. (Eds.) *Migration and media: Discourses about identities in crisis.* (pp. 63-90). John Benjamins.

Taylor, C. (2009). The representation of immigrants in the Italian press. *CirCap Occasional Papers 21*. Siena: University of Siena.

Zanettin, F. (2014). Corpora in translation. In House, J. (ed.) *Translation: A multidisciplinary approach* (pp. 178-199). Palgrave Macmillan.

Investigating the Role of Entrenchment in Source Language Interference during Simultaneous Interpreting

Jūratė Žukauskaitė

University of Agder, Norway

KEYWORDS: source language interference, cognitive linguistics, entrenchment, simultaneous interpreting, corpus study

Source language interference (SLI) is a ubiquitous phenomenon in interpreting, but the cognitive processes that underlie it have not been extensively studied. This research aims to fill this gap and test a cognitive linguistic/psycholinguistic account of lexical SLI based on the Gravitational Pull hypothesis (Halverson, 2003) and the cognitive linguistic concept of entrenchment. According to this concept, the strength of cognitive representations and the connections between them is increased by repeated (co-)activation (Langacker, 1987). In the proposed account, the bilingual semantic network, specifically the entrenched connections between word forms and their senses, is suggested as the locus of SLI. The hypothesis is that having to translate a less entrenched sense of a polysemous word into the target language increases the likelihood of SLI.

To test this hypothesis, a study will be conducted using the EPIC UdS simultaneous interpreting corpus (Przybyl et al., 2022) to compare the performance of English-German interpreters when translating words with different entrenchment patterns. The words used in the study will be determined through an entrenchment study employing a methodological approach that combines three different methods to access the directly unobservable cognitive phenomenon of entrenchment and ensure that the resulting description of the crosslinguistic entrenchment patterns is as psychologically realistic as possible. Entrenchment of the senses will be operationalized as frequency of their use, following Halverson (2017). Cross-linguistic relationships for a sample of selected polysemous words will be established through the utilization of an English language corpus (counting the frequency of the senses), as well as sentence generation and translation tasks involving German native speakers enrolled in a Master's program in English-German translation. Both tasks will yield data regarding the primary sense employed/translated and the frequency of sense usage. The entrenchment study will result in two subsets of polysemous words with two different entrenchment patterns: words with one highly entrenched sense and at least one non-entrenched sense, and words with at least two senses that are entrenched at a similar level. An entrenchment score will be computed for the senses, derived from the results of the corpus analysis, the sentence generation task, and the translation task.

The resulting subsets of polysemous words will be used in the interpreting corpus study to compare the incidence of SLI when interpreting polysemous words with different entrenchment patterns. The incidence of SLI will be measured by counting the frequencies of SLI-related errors (selecting an inappropriate translation equivalent) and disfluencies that can be interpreted as signs of SLI (pauses, self-corrections, mispronunciations, and false starts). It is hypothesized that the incidence of SLI will depend on the entrenchment pattern, particularly on the entrenchment score assigned to the senses in the entrenchment study: the lower the entrenchment score (i.e., the less entrenched a sense), the higher the incidence of SLI should be when rendering the sense into the target language.

References

Halverson, S.L. (2003). The cognitive basis of translation universals. *Target-international Journal of Translation Studies, 15*, 197–241.

Halverson, S. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical Translation Studies* (pp. 9–46). De Gruyter Mouton.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Vol. Vol. 1*. Stanford University Press.

Przybyl, H., Lapshinova-Koltunski, E., Menzel, K., Fischer, S., & Teich, E. (2022). EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1193–1200.

MateCat performance as a basis for devising quality assessment metrics for legal texts' automated translations

Edyta Źrałka

University of Silesia, Poland

KEYWORDS: translation quality assessment, machine translation (MT) corpora, MateCat performance, post-editing, specialised translation

Juliane House's idea of Translation Quality Assessment (1997), imported and further discussed by Basil Hatim (1998) and others, appeared particularly applicable in the evaluation and post-editing of machine translations (MTs). Due to the awareness of the reduced credibility of such tools as MT engines, including those gratuitous and commonly used, e.g. Google Translate (GT) or DeepL, the necessity of quality control through automatic or human metrics and post-editing such translations for their professional use occurred. Some aiding instruments for the evaluation procedure had to be introduced (metrics) to facilitate post-editing and make the process of MT valuable and sensible, subject to specified rules. The idea was not only to invent the criteria according to which metrics could be designed and the evaluation performed. To record the proofread outcomes somehow and make them repetitive constituted a challenge. In such a case, the process of evaluation and post-editing could be more effective and would transfer to better translation quality. A valuable contribution to that need was the creation of MateCat (Machine Translation Enhanced Computer Assisted Translation) tool, which is a combination of MT engine and CAT tool, enabling a user to translate automatically and to edit MT results for better and better outcomes based on translation memories (TMs) and terminology databases. The outcomes are dependent on the parallel corpora the tool is equipped with. However, what is often left random with GT and DeepL., even if based on authentic texts, in the case of MateCat can be systematised due to TMs, and the corpora serve as a source of linguistic data to improve the quality of translation in subsequent translations. Through the suggested corrections available, the categories of typical errors can be better recognised and serve to establish the criteria for creating quality assessment metrics to evaluate and post-edit also GT or DeepL translations.

Despite the improved quality of MateCat performance compared with GT or DeepL, which was taken for granted, all instruments need quality assessment and post-editing to serve professional aims and enable translators to use them in order to automatically translate specialised texts, sometimes as challenging as, e.g. legal texts, and raise productivity. Metrics are created based on the particular needs dictated by translated texts. Even if there are other existing metrics, the new ones are created to be specially adjusted to particular needs of an agency, translator etc. The MateCat availabilities, then, can serve as a source of essential data to design both manual and automatic metrics and systematically improve MT quality of legal texts. Having an outcome of MateCat and a manual metric for evaluation, a translator can easily post-edit his or her translations quickly and make them high quality, reducing time and effort.

The observation of categories of errors corrected based on the content of MateCat corpora can be a valuable suggestion for the criteria selected for creating metrics. This paper aims to find them using MateCat outcomes. When discovered and incorporated in the process of creating the metric for legal texts' translation quality assessment, they could serve as the basis upon which a translator can develop his or her MT quality assessment rules and routines to post-edit automatically translated specialised texts in an organised and effective way.

The procedure of creating the metric is based on translating legal texts from English into Polish through MateCat and observing sources of most typical errors. The texts chosen for carrying out the study are four contracts of different types (Open-ended Employment Agreement, Mandate Agreement, and Donation Agreement), each comprising between 1,000 and 2,500 words (1,452, 1,123, 2,393, 1,365, respectively). The type of legal documents was selected based on its functional importance in legal translation for a typical client (the agreements are documents which freelance translators often deal with). Research questions that come to mind comprise those which refer to the quality of MT of the sample texts (to what

extent is the MateCat translation acceptable, may it be left non-post-edited, what types of errors are the most common, what ideas to improve the translation quality can be observed with the tool used), and the effectiveness of adding CAT tools opportunities (how much CAT tools' TMs, databases and corpora help to achieve an acceptable quality of translation at the level of MateCat performance and what elements can be included in the metric designed to improve quality of GT and DeepL translation if a translator decides to use MT tools instead).

To prepare the criteria for the metric, possible errors in the MateCat translations will be classified based on a comparative study of the source text (ST) in English and the target text (TT) in Polish generated by the tool in the first search for the equivalents. Possible categories of errors will constitute the basis for establishing criteria for creating the metric, which will be manual and aimed at improving both the quality and pace of post-editing.

Machine translation and the theoretical aspects of the MateCat will be discussed based on Kenny (2018), Federico et al. (2014), Nurjanah (2021), and others.

References

Domínguez Mora, E. & Ana Ibañez Moreno, U. (2022, May. 28). Comparative analysis of Google Translate and DeepL in Spanish-English literary translation: The case of collocations in Don Quixote [Paper presentation]. Positive impacts of language technology: TISLID 22, Madrid, Spain.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., ... & Germann, U. (2014, August). The MateCat tool. In *COLING (Demos)* (pp. 129-132).

Forcada, M. L. (2010). Machine translation today. *Handbook of translation studies*, *1*, 215-223.

Han, L., Jones, G. J., & Smeaton, A. F. (2021). Translation quality assessment: A brief survey on manual and automatic methods. *arXiv preprint arXiv:2105.03311*.

Hatim, B. (1998). Translation quality assessment: Setting and maintaining a trend. *The Translator*, *4*(1), 91-100.

House, J. (1997). Translation quality assessment: A model revisited. Gunter Narr Verlag.

House, J. (2015). *Translation quality assessment*. Routledge.

Jopek-Bosiacka, A. (2006). *Przekład prawny i sądowy*. Warszawa: Wydawnictwo Naukowe PWN.

Kenny, D. (2018). Machine translation. In *The Routledge handbook of translation and philosophy* (pp. 428-445). Routledge.

Kodeks zawodowy tłumacza przysięgłego. (2019). Warszawa: TEPiS. https://tepis.org.pl/wp-content/uploads/Kodeks-zawodowy-t%C5%82umacza-przysie%CC%A8g%C5%82ego.pdf

Maučec, M. S., & Donaj, G. (2019). Machine translation and the evaluation of its quality. *Recent trends in computational intelligence*, 143-162.

Mellinkoff, D. (2004). *The language of the law*. Eugene: Resource Publications.

Nitzke, J., Hansen-Schirra, S., & Canfora, C. (2019). Risk management and post-editing competence. *The Journal of Specialised Translation*, *31*, 239-259.

Nurjanah, R. L. (2021, November). Matecat. com: Why is It Worth to Try?. In ELTLT 2020: Proceedings of the 9th UNNES Virtual International Conference on English Language Teaching, Literature, and Translation, ELTLT 2020, 14-15 November 2020, Semarang, Indonesia (p. 49). European Alliance for Innovation.

Sadollah, A. & Sinha, T. S. (Eds.). (2020). *Recent trends in computational intelligence*. BoD–Books on Demand.

Tomás, J., Mas, J. À., & Casacuberta, F. (2003, April). A quantitative method for machine translation evaluation. In Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable? (pp. 27-34).

Source texts:

https://www.boardfoundation.org/board-guidelines/mandate-agreement.pdf

https://www.macedovitorino.com/xms/files/Why_Portugal/Direito_laboral/Open-Ended_Employment_Contract.pdf

https://illinois.fntic.com/FNTRSIllinois/media/media/Online%20Documents/Sales%20Contract/RealEstateContractFormC.pdf

https://broadview-il.gov/wp-content/uploads/2019/02/CHICAGO-225946-v3-Donation_Agreement.pdf

Bionotes

Mohammad Aboomar

Mohammad Aboomar is a PhD researcher at Dublin City University, Ireland and an awardee of the Irish Research Council postgraduate scholarship. He is currently researching the effect of culture on the contemporary translation of evolutionary biology into Arabic. His research interests include scientific translation, translation history, terminology, indirect translation, and corpus linguistics.

Francesca Antonioli

Francesca Antonioli is a student in the Master of Advanced Studies in Linguistics at Ghent University. She previously obtained a Master's degree in Specialized Translation at the University of Trieste. Currently, her research interests mainly concern translation and the (cognitive) factors that condition translators in linguistic decision-making.

Amjed Al-Rickaby

Amjed Al-Rickaby is a PhD researcher at the University of Aberdeen, UK. He specialises in contrastive linguistics. Al-Rickaby's main research interests include corpora design, critical discourse analysis of English and Arabic data, contrastive pragmatics, effect of native culture on second language production. He has published several papers and books in these research fields.

Reem Alfuraih

Reem Alfuraih (Princess Nourah bint Abdulrahman University), A lecturer at Princess Nourah bint Abdulrahman University, College of Languages, Applied Linguistics Department. PNU Research Services Center Director and the managing editor of 'Critical Studies in Languages and Literature' journal. The principal invistigator of the The undergraduate learner translator corpus (ULTC) project.

Aladdin Al Zahran

Dr Aladdin Al Zahran received his BA in English Language & Literature from Aleppo University, Syria, (1999) and MA in Arabic/English Translation & Interpreting (2004) and PhD in Interpreting Studies (2007) from Salford University, UK. Dr Al Zahran is currently an Assistant Professor of Interpreting & Translation Studies at Sohar University (Oman), where he has served as the Programme Coordinator of the BA in English Language and Translation since September 2018. He has taught courses across the spectrum in translation, interpreting, language and linguistics. His research interests include intercultural mediation, the interpreter/translator's role, interpreting strategies, and corpus-based interpreting/translation studies. He is also a freelance interpreter, translator, reviser, external examiner and translation consultant.

Khatuna Beridze

Dr. Khatuna Beridze (Associate Professor at Batumi Shota Rustaveli State University)

Her research interests include Translation Studies, Literary Translation, Corpus Linguistics, Postcolonialism, Translators Training Methodology, Terminology. Kh. Beridze is the founding head of the Translation and Interdisciplinary Research Center at BSU, collaborating with international and local scientific units. In 2009, 2011 and 2012, she was a visiting scholar at Illinois State University and the Harriman Institute, Columbia University. Khatuna Beridze is the author, co-author and editor of over 50 scientific papers and publications in the fields of postcolonialism, corpus linguistics and literary translation.

The main work in the field of postcolonial translation is the monograph: "Translation and the Theory of Postcolonialism, Translation Imperialism" (in Georgian, 2018). In the field of corpus linguistics, she authored the BSU Parallel Research Corpus © www.corpus.bsu.edu.ge, a Parallel Corpus of Political Texts and Terminology; To the field of Translation Studies she devoted a textbook "Translation Studies" (in Georgian, 2018, BSU).

She is a practicing translator and conference-interpreter and has recently developed international MA programme: Translation and Conference Interpreting.

Silvia Bernardini

Silvia Bernardini is professor of English linguistics and translation studies at the Department of Interpreting and Translation of the University of Bologna, Italy, where she teaches translation from English into Italian and corpus linguistics. She holds an MPhil in applied linguistics from the University of Cambridge and a PhD in translation studies from the University of Middlesex. She has published widely on corpus creation and use in translator education and for translation practice and research. Her research interests include the creation of corpus resources, the investigation of constrained language, and didactic innovation in translation teaching through service learning and research-based teaching.

Łucja Biel

Łucja Biel is a Professor of Linguistics and Translation Studies and Head of EUMultiLingua research group in the Institute of Applied Linguistics, University of Warsaw, Poland. She is the editor-in-chief of the *Journal of Specialised Translation* and secretary of Council of Editors of Translation and Interpreting Studies for Open Science. She has published extensively on EU/legal translation, legal terminology, translator training and corpus linguistics. She is also a sworn translator with over 25 years of professional experience.

Romane Bodart

Romane Bodart is a teaching and research assistant at UCLouvain, where she teaches legal translation to first- and second-year master's students. Her PhD thesis deals with machine translation post-editing (MTPE) in translator education. Her current work focuses on the comparison of post-editing and human translation quality.

Antonina Bondarenko

Antonina Bondarenko recently defended a PhD dissertation studying verbless sentences from a contrastive corpus perspective at the Université Paris Cité and INALCO. She is currently a postdoctoral fellow with CNRS (SeDyL) and an associate research fellow at Université Paris Cité (Clillac-Arp), as well as part of several projects targeting non-canonical questions, information structure, and, more recently, monologic discourse (SurQue, Marco, TransQuest). Her main research focus has been on developing a

methodological framework that combines contrastive, enunciative and corpus linguistics in order to gain semantico-pragmatic insights into the verbless phenomenon and elaborating a model of the sentence capable of accounting for the structures.

Lynne Bowker

Lynne Bowker is Full Professor at the University of Ottawa in Canada, where she holds a cross-appointment between the School of Translation and Interpretation and the School of Information Studies. She teaches and conducts research at the intersection of language and technology. During the 2022/2023 academic year, she is a NAWA Visiting Researcher in the Scholarly Communication Research Group at the Adam Mickiewicz University in Poland, where she is working on a project about plain language and machine translation. She is the author of several books, including Computer-Aided Translation Technology (University of Ottawa Press, 2002), Working with Specialized Language: A Practical Guide to Using Corpora (Routledge, 2002), Machine Translation and Global Research (Emerald, 2019) and De-mystifying Translation (Routledge, 2023).

Anastasia Buturlakina

Anastasia Buturlakina is PhD student in translation studies at Université Paris Cité, France. For her Master's thesis she developed a semi-automated methodology to detect humorous prosodic clashes in the corpus of TED Talks transcripts. Her PhD research is focused on translation of humorous linguistic incongruities between English and French in TED Talks. Her research interests are humour translation, semantic prosody and corpora.

Cheng Yu Shen

Cheng Yu Shen is a PhD student of Translation and Language Sciences at Universitat Pompeu Fabra. Her research interests are primarily based on interactions between language and ideology in political discourse delivered in Mandarin. She will be working on the how discursive constructions of Taiwanese citizenship have changed diachronically in political discourses with the application of systemic functional grammar.

Regina Dasaeva

Regina Dasaeva is a master's degree student in Philology in Kazan Federal University as well as a graduate of the same University in Economics. She works as a leading economist in Republican Clinical Hospital. For her research she uses the LancsBox programme as she completed the Lancaster's course on FutureLearn

Joke Daems

Joke Daems is assistant professor human-computer interaction in empirical translation & interpreting studies at Ghent University. They conduct research as a member of EQTIS (Empirical and Quantitative Translation and Interpreting Studies) and collaborate closely with the Language and Translation Technology Team (LT^3). Their research interests include (machine) translation, post-editing, translatability, human-computer interaction, and literary translation and style.

Gert De Sutter

Gert De Sutter is currently working as Professor in Translation Studies and Dutch Linguistics in the Department of Translation, Interpreting and Communication at Ghent University. He is vice-head of the Department, programme director of the Master of Translation and coordinator of the EQTIS research unit. His research can be broadly characterised as empirical research of linguistic variation in translated and non-translated texts. The overall aim of his research is to get a better insight into the complex interaction of social norms and bilingual cognition in translators. To that end, he draws on descriptive, methodological and theoretical insights from usage-based linguistics, sociolinguistics and psycholinguistics.

Bart Defrancq

Bart Defrancq is an associate professor at Ghent University and course leader of the interpreter training programmes. His research focuses on simultaneous interpreting and police interpreting, mostly from a corpus-based perspective

Vera Demberg

Vera Demberg is a W3 professor at Saarland University, affiliated with the Department of Computer Science and the Department of Language Science and Technology. Her research interests include cognitive models of human language understanding, natural language generation, experimental and computational discourse and pragmatics and multimodality."

Khatuna Diasamidze

Dr. Khatuna Diasamidze is a senior lecturer and has extensive working experience as a teacher of translation. She is a practicing translator. Dr. Diasamidze was awarded Doctoral degree at Batumi Shota Rustaveli State University. Her Ph.D. thesis supervisor was Dr. Assoc. Prof. Beridze.

Adriano Ferraresi

Adriano Ferraresi is associate professor in English language and translation at the University of Bologna, where he teaches MA modules in translation technologies. His research focuses on phraseology and terminology in native, translated and lingua franca varieties of English, mostly through quantitative, corpus-based methods. He also contributes to various collaborative corpus building initiatives, including recently the EPTIC and CHEU-Lex projects.

Patrícia Helena Freitag

Patrícia Helena Freitag is a doctoral student at Universidade Federal do Rio Grande do Sul (UFRGS), Brazil. Her current research is on the similarities and differences of collocational patterns in texts translated into Portuguese versus texts originally written in Portuguese. She also holds a Master's degree in Psycholinguistics, a Bachelor's degree in Translation and an MBA in Project Management. As a translator, she has some published translations (e.g. "Heróis do Xadrez Clássico," "Gigantes do Xadrez Agressivo," and "Crianças, Espaços, Relações: Como Projetar Ambientes para a Educação Infantil"), but

her main area of activity is with language service providers. She is currently a member of Professor Rozane Rebechi's project "Testagem do Dicionário Gastronômico Português-Inglês.""

Jonas Freiwald

Jonas Freiwald is a teaching and research assistant at the Department of English Linguistics at RWTH Aachen University, Germany, which is where he received his PhD in English linguistics. He is a member of the TRICKLET project, a multi-method translation studies project on translation process and translation product data. His research interests include contrastive linguistics, translation studies and corpus linguistics and he has a background in functional grammar analysis.

Sílvia Gabarró-López

Sílvia Gabarró-López is a postdoctoral researcher at Pompeu Fabra University (Barcelona) and a scientific collaborator at University of Namur (Belgium), from which she received her PhD in 2017. Her thesis is the first to study discourse markers from a comparative corpus-based perspective in two sign languages. Before holding her current positions, she lectured on Linguistics and Deaf Culture at University Saint-Louis-Brussels and at the Catholic University of Louvain (Belgium). She also worked as a postdoctoral researcher at Stockholm University (Sweden). She is mainly interested in visual and tactile sign languages, discourse analysis, multimodality, contrastive linguistics and interpreting.

Federico Garcea

Federico Garcea is an adjunct professor at the Department of Interpreting and Translation of the University of Bologna, teaching students of the Translation and Technology master's degree. Prior to joining the University, he was responsible for the development of Machine Translation service at Microsoft Research. His research focuses on the assessment and value of translation technologies, and their application in difficult domains like artistic and cultural content, preservation and transmission of cultural heritage.

Łukasz Grabowski

Prof. Łukasz Grabowski is Professor in Linguistics at the University of Opole, Poland. His research interests are corpus linguistics, phraseology, formulaic language, translation studies, lexicography, and ESP.

Sandra L. Halverson

Sandra L. Halverson is employed at Agder University in Norway. Her research has centered on questions related to various areas of Translation Studies and Cognitive Linguistics, and she has published both empirical and theoretical/conceptual work. An overarching concern is the integration of insights from Cognitive Linguistics into Translation Studies, and she is currently working on hypotheses linking translational choices to specifics of cognitive representation and processing. Other long-term research interests are the epistemology of Translation Studies and research methodology. She was appointed CETRA Chair Professor for 2018 and is an external associate of the MC2 Lab and a member of the TREC and INTERACT networks.

Daniel Henkel

Daniel Henkel is Associate Professor of Linguistics and Translation at Université Paris 8, where he teaches Computer Assisted Translation, Terminology and Comparative Syntax. His research focuses on translation between English, French and Italian using quantitative methods to compare corpora of translated and original texts. He has also translated numerous articles published in international scientific journals.

Alice Heylens

Alice Heylens is a PhD student in UNamur and UCLouvain under the supervision of Laurence Meurant and Marie-Aude Lefer. She is the first and only, so far, graduated translator in French Belgian Sign Language (LSFB) and English. After working three years as a translator at the LSFB-Lab in UNamur, she started a PhD which aims to characterize translations from signed to spoken languages.

Hung-Hsin Hsu

Currently, I am a PhD student of a joint doctoral degree of Université catholique de Louvain (Belgium) and National Chengchi University (Taiwan). My primary research interests are the syntax of Mandarin and French, second language acquisition, and historical linguistics. The languages studied include Mandarin, Taiwan Southern Min, English, and French. The works that I have presented in conferences cover topics in syntax, corpus linguistics, cognitive grammar, and second language acquisition. In addition to conference papers, I also have one peer-reviewed paper published in the Journal of Chinese Language Teaching (THCI, ACI). This study focuses on the question particle ne in Mandarin. Aside from personal research interests, I also work with one of my supervisors, Prof. Her One-Soon, on numeral classifier languages in Africa.

In addition to research, teaching is also what I am interested in. I was able to gained valuable experience when I served as a teaching assistant for a course called The Humanities and Sciences of Language. And I was honored to be selected as the Distinguished and Outstanding Teaching Assistant in 2020 and in 2021 respectively.

Ilmari Ivaska

Ilmari Ivaska is assistant professor of Finnish at the School of Languages and Translation Studies of the University of Turku, Finland. His research focuses on data-driven corpus methodologies and their applications in studying typical tendencies in texts written by L2 language users and, more recently, translated texts written by L1 language users, when compared to non-translated texts written by L1 language users. The investigation of new ways to study linguistic divergence between different modes and varieties of language use lie at the core of Ivaska’s research interests.

Laura Ivaska

Dr. Laura Ivaska is currently university teacher of English at the University of Turku. She is on leave from the research project “Traces of translation in the archives” at the Finnish Literature Society (SKS). Her research interests include indirect translation, translation history, genetic translation studies, theory of text, and corpus studies. She is also co-coordinator of the international IndirecTrans network and editor of MikaEL (the Electronic Journal of the KäTu Symposium on Translation and Interpreting Studies)."

Anna Jelec

Anna Jelec is an assistant professor at Adam Mickiewicz University in Poznan (Poland) and a conference interpreter and interpreter trainer with over 10 years of experience in the field. She is interested in cross-disciplinary research, in particular cognitive processes in interpreting and the applications of research results to interpreter training.

Alina Karakanta

Alina Karakanta is a machine translation researcher and professional translator. She obtained her PhD in Information and Communication Technology from the University of Trento and Fondazione Bruno Kessler on the topic of automatic subtitling. Her research interests cross the interdisciplinary paths between machine translation, speech technologies, audiovisual translation, corpus-based translation & interpreting studies and evaluation methodologies. She has given several invited lectures and talks on speech translation and its impact on translation workflows. She holds a MSc in Computational Linguistics from Saarland University and BAs in Translation Studies and Interpreting Studies from the Ionian University.

Eva Klüber

Eva Klüber is a PhD student at Heidelberg University. Her research is focused on investigating semantic transfer and cognitive load in interpreting, and applies corpus-based methodologies."

Janina Kröger

Janina Kröger has a master's degree in media linguistics and is a doctoral student at the University of Hildesheim. She is part of the interdisciplinary stipend programme "Chronic Diseases and Health Literacy" ("Chronische Erkrankung und Gesundheitskompetenz", ChEG), funded by the Robert-Bosch-Stiftung at the Hanover Medical School (Medizinische Hochschule Hannover). She is also part of the research group "Accessible Medical Communication" at the University of Hildesheim. Janina Kröger works at the Research Centre for Easy Language.

Kerstin Kunz

Kerstin Kunz is a professor for English Translation studies at the Institute for Translation and Interpreting at Heidelberg University. She applies quantitative and qualitative methods to explore language contrast, register variation, translation and interpreting strategies on the level of lexicogrammar and discourse.

Michiel Kusé

Michiel Kusé is currently working as a PhD candidate at Ghent University's Department of Translation, Interpreting, and Communication (VTC), where he studies the occurrence of default translation patterns during the translation and interpreting processes, and how such patterns are influenced by experience and training on the one hand, and by linguistic distance between source and target language on the other. He holds degrees in Translation and Literary Studies from the Catholic University of Leuven, and is currently working towards a degree in Psychology (Ghent University).

Natalie Kübler

Natalie Kübler is Professor in English linguistics and translation studies. She directed a European project on a multilingual multidirectional translation learner corpus (MeLLANGE). She's been the head of her research team since 2014. Her research interests are corpora and specialized translation, bilingual terminology, bilingual specialized phraseology, and specialized MT. "

Lisa Marie Lang

Lisa Marie Lang did the BA in Slavic Studies, then the MA in Translation Studies (with focus on Russian). She spend some semestres abroad (Russia and Belgium), attented a sommer school abroad (Siberia) and has work experience as journalist (Kazakhstan) and as a freelance translator (Russian and French). Now she is a PhD candidate in linguistics.

Ekaterina Lapshinova-Koltunski

Ekaterina Lapshinova-Koltunski is a Full Professor in Multilingual Technical Specialised Communication at the Stiftung University of Hildesheim. Her areas of interest include linguistic variation in multilingual contexts, with a focus on variation in translation. She has been working with corpus-based and corpus-driven methods using techniques derived from data mining.

Marie-Aude Lefer

Marie-Aude Lefer is Associate Professor of Translation Studies and English-French translation at UCLouvain, where she acts as Head of the Louvain School of Translation and Interpreting. Her current research interests include multilingual corpus compilation, corpus use in translator education, learner translation corpus research, error annotation, quality evaluation and machine translation post-editing.

Agnieszka Leńko-Szymańska

Agnieszka Leńko-Szymańska is Assistant Professor at the Institute of Applied Linguistics, University of Warsaw, Poland. Her research interests are primarily in second language acquisition and corpus linguistics, as well as in data-driven learning and teaching. She has authored a number of papers on corpus-based explorations of L2 vocabulary and formulaic language and on training language teachers in corpus methods. Her most recent publications include a monograph Defining and Assessing Lexical Proficiency (Routledge, 2020) and a chapter 'Training Teachers and Learners to Use Corpora' in The Routledge Handbook of Corpora and English Language Teaching and Learning, edited by R. R. Jablonkai and E. Csomay (Routledge, 2022). She also co-edited two monographs: Multiple Affordances of Language Corpora for Data-driven Learning and Complexity, Accuracy and Fluency in Learner Corpus Research, both published by John Benjamins in 2015 and 2022 respectively.

Azad Mammadov

Azad Mammadov (PhD 1990, Doctor of Sciences 2004) is Chair Professor of General Linguistics at Azerbaijan University of Languages, Baku, Azerbaijan. Much of his work has been devoted to the study of text and discourse from a pragmatic, cognitive and sociolinguistic perspective. His publications in this field include Studies in Text and Discourse (2017) and Repetition in Discourse (2019, with Misgar

Mammadov and Jamila Rasulova) as well as several scholarly papers. (International Journal of Sociology of Language, Lodz Papers in Pragmatics, Asia-Pacific Translation and Intercultural Studies Journal, International Review of Pragmatics and English Today among others) He is co-editor of Analyzing Media Discourse: Traditional and New, a Cambridge Scholars Publishing Volume and guest co-editor of Exploring Dialogical Discourse: Pragmatics and Cognition, Special Issue of International Review of Pragmatics.

Nathália Oliva Marcon

Nathália Oliva Marcon is an undergraduate student at Universidade Federal do Rio Grande do Sul (UFRGS) — Brazil, currently pursuing a Bachelor's degree in Linguistics, majoring in Portuguese/English. She is a project team member in Professor Rozane Rebechi's project "Testagem do Dicionário Gastronômico Português-Inglês" and in Professor Marcia Moura da Silva's project "ABREVITRAD: tradução de abreviaturas e acrônimos de termos médicos no par linguístico Português-Inglês".

Lorenzo Mastropierro

Dr Lorenzo Mastropierro is lecturer in English Language and Translation at the University of Insubria, Italy. His research interests are corpus linguistics, stylistics, and translation studies.

Judyta Mężyk

Judyta Mężyk is a PhD candidate in linguistics at the University of Silesia in Katowice and Paris-East Créteil University in Paris. She is also an assistant (which is a research-teaching position) at the University of Silesia and the Tischner European University in Cracow. She is the author of a few articles and two book chapters on linguistics and translation. Her research interests involve particularly audiovisual translation, corpus linguistics, formulaic language, and interpreting as a non-professional practice.

Zoë Miljanović

Zoë Miljanović (M.A. in Linguistics and Literary Studies) is a teaching and research assistant at the Institute for English and American Studies at RWTH Aachen University in Germany and part of the TRICKLET (Translation Research in Corpora, Keystroke Logging and Eye Tracking) project team. In her dissertation project, she is working on a bi-directional corpus-based investigation of literary translation over time for the language pair English-German. Her research interests lie in the fields of descriptive translation studies, translation history, stylistics, and systemic functional grammar.

Olga Nádvorníková

Olga Nádvorníková is senior lecturer at the Department of Romance Studies of the Faculty of Arts of Charles University in Prague. She has published widely in the field of corpus-based contrastive and translation studies, comparing Czech, French, English and recently also Polish (research into the so-called translation universals, especially explicitation and normalization, studies about shifts in the segmentation of sentences in translation, contrastive analysis of Czech, French and Polish converbs, etc.). Since 2022, she is member of the permanent organizing committee of the Translation in Transition conference series.

Stella Neumann

Stella Neumann is a full professor of English Linguistics at RWTH Aachen University, Germany. She holds a degree in translation from the University of Bonn, Germany and received a PhD in translation studies and a habilitation in English linguistics from Saarland University, Saarbrücken, Germany. Her research interests include the combination of corpus-linguistic and experimental methods for the empirical modelling of translation as well as quantitative register analysis across varieties of English and languages.

Maja Miličević Petrović

Maja Miličević Petrović is an associate professor in the Department of Interpreting and Translation of the University of Bologna at Forlì, teaching general linguistics, second language acquisition and language data analysis. She is interested in graded language phenomena, especially those at the interface between morphosyntax and lexical semantics, and in different types of bilingual contexts, especially comparisons between translation, second language acquisition and first language attrition.

Christina Pollkläsener

Christina Pollkläsener is a master's student at Saarland University. She has just submitted her master's thesis on the comparison of discourse connectives in translated in interpreted texts.

Thomas Prinzie

Thomas Prinzie grew up in a multilingual family in Belgium. He has also lived in the US and Canada, where he has pursued a range of interests including music, evolutionary biology, and Buddhist meditation. After returning to Belgium in 2019, Thomas completed a Master in linguistics at UCLouvain. He is now a PhD student under the supervision of Marie-Aude Lefer at UCLouvain's Centre for English Corpus Linguistics. His project concerns the roles of cognition and expertise in the French translation of English noun sequences. Thomas has always been fascinated by the partially overlapping patterns of meaning and form in different languages. His research interests include corpus-based contrastive linguistics and translation studies, cognitive linguistics, bilingual mental representation, and intensification.

Heike Przybyl

Heike Przybyl is a researcher at Saarland University. She investigates linguistic characteristics of interpretese in a corpus-based approach. With a focus on simplification, she aims to link cognitive processing in simultaneous interpreting with linguistic properties found in the interpreting product.

Heike was involved in the corpus building project EPIC-UdS and is a member of the Collaborative Research Centre 1102 on Information density and linguistic encoding, project B7 on Translation as rational communication.

Heike is also a trained conference interpreter and occasionally still enjoys working as interpreter.

Rozane Rebechi

Rozane Rebechi is a professor and researcher at the Federal University of Rio Grande do Sul, where she is also head of the Modern Languages Department. She holds a Master and a Ph.D. degree in English

Language and Literature from the University of São Paulo (Brazil). Her main areas of research are Translation, Terminology, and Discourse, to which she applies Corpus Linguistics as methodology. She was chair of the Brazilian Association of Researchers in Translation (ABRAPT) through 2020-2022 and is Associated Partner and member of the management board to the European Masters in Technology for Translation and Interpreting (EM TTI).

Matt Riemland

Matt Riemland is a PhD candidate in Dublin City University's School of Applied Language and Intercultural Studies. He is supervised by Dorothy Kenny (Dublin City University) and James Hadley (Trinity College Dublin). His research interests include corpus-based translation studies, machine translation, translation in crisis and development settings, legal translation, and the influence of power dynamics on translation products and processes.

Natalia Rodriguez Blanco

Natalia Rodriguez Blanco is a PhD student in Interpreting, Translation, and Intercultural Studies at the University of Bologna. She is a linguist, a professional interpreter, and translator (ES/FR/EN), whose research interests derive from her own experience working for national and international institutions. Her doctoral research project revolves around the areas of translation, multilingual and global news production, sociocultural representation, and discourse analysis. Her research intersects translation and corpus studies, often accompanied by fieldwork.

Alexandr Rosen

Alexandr Rosen is a researcher and lecturer at the Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University in Prague. His main professional interest is linguistic theory and corpus linguistics. He has been working on parallel and learner corpora and their annotation, taxonomy of linguistic categories and linguistic theories. At the Institute of the Czech National Corpus, he leads a team building the InterCorp parallel corpus.

Rana Roshdy

Rana Roshdy is a Ph.D. Candidate at Dublin City University in Ireland. She is working on a corpus-based project that investigates the translation of Islamic law into English. Her Ph.D. research is under the supervision of Professor Dorothy Kenny and funded by the Irish Research Council and DCU's School of Applied Language and Intercultural Studies.

Rana has a B.A. in English Literature and Linguistics, an M.A. in the Cultural Politics of Translation, and a diploma in UN and legal translation. As a practitioner, Rana specializes in economic translation, having worked as a Senior Translator/Editor at Naeem Holding, Egypt.

She had her first academic publication in Alif-38: Translation and the Production of Knowledge(s), edited by Professor Mona Baker. Her research interests focus on specialised translation (legal and financial), corpus linguistics, translation technology, Islamic law, and finance."

Paulina Rozkrut

Paulina Rozkrutis a PhD student at the Doctoral School of Languages and Literatures of Adam Mickiewicz University in Poznań, Poland. She finished her Master's studies in Polish<>English conference interpreting and is now working on her PhD project that will investigate the use of automatic speech recognition (ASR) by trainee and professional interpreters. Her research interests include, among others, computer-assisted interpreting, human-computer interaction, and corpus linguistics.

Anna Setkowicz-Ryszka

Anna Setkowicz-Ryszka (Doctoral School of Humanities, University of Łódź, Poland), Translator with 25 years' experience, specializing in legal, financial and academic texts, sworn translator of English in Poland, legal translation trainer. Long-standing cooperation, including also revision and post-editing, with legal publishing houses and law firms. PhD student at the Doctoral School of Humanities, University of Lodz, Poland. Research interests: legal translation pedagogy; contract translation with special emphasis on boilerplate clauses; plain legal language; academic legal language; machine translation post-editing in the field of law; translation process research; translator expertise.

Merel Scholman

Merel Scholman is an assistant professor at Utrecht University. She is interested in cognitive models of language understanding, focusing on discourse coherence in particular. She uses a combination of corpus-based investigations, crowdsourced studies, and reading time experiments to investigate the processing of discourse.

Pang Shuangzi

Pang Shuangzi is a tenured associate professor at the School of Foreign Languages, Shanghai Jiao Tong University. She holds a PhD in corpus-based translation studies from Beijing Foreign Studies University, China. Her academic interests lie in corpus-based translation studies, language contact, and the relation between translation and language change. She has achieved 3 projects awarded by The National Social Science Fund of China as a principal investigator. She has More than 10 articles published in top periodicals in linguistics and translation studies both in China and in the west, such as Target: International Journal of Translation Studies, Foreign language teaching and research, Chinese Translators Journal and Journal of Foreign Languages. Some of them have been cited and reprinted in Replicated Journals of Renmin University of China and China Social Science Network.

Cinzia Spinzi

Cinzia Spinzi is Associate Professor at the University of Bergamo. She holds a PhD in English for Specific Purposes, a Master's in Translation Studies from the University of Birmingham, and a Research Fellowship from the City University of London.

Her research activity and major publications lie in Language Mediation, Cross-cultural communication and Translation, with a particular focus on the translation of tourism, accessibility and metaphors.

She is member of the Research Centre on Languages for Specific Purposes (Cerlis) and of the EU funded Project TTRAILs on teaching and training in the field of Language for Specific Purposes.

She is co-editor of an international journal Cultus: the journal of Intercultural Mediation and Communication. (www.cultusjournal.com).

orcid.org/0000-0003-3267-6905"

Sofie Verliefde

Sofie Verliefde is a postdoc researcher at Ghent University. She recently defended her PhD on the interaction of interpreting and drafting in interpreter-mediated police interviews.

Jelena Vranjes

Jelena Vranjes is post-doctoral researcher at the Department Translation, Interpreting and Communication at Ghent University. Her research interests include interpreting, conversation analysis, multimodality in human interaction and the impact of context and technology on the perception and use of embodied resources. She uses advanced empirical methods, such as eye-tracking, in order to gain a better understanding of the multimodal dynamics of interpersonal talk.

Katarzyna Wasilewska

Katarzyna Wasilewska is Assistant Professor at the Institute of Applied Linguistics, University of Warsaw, Poland. She has recently defended her Ph.D. dissertation entitled ""Administrative reports: a corpus study of the genre in the EU and Polish national settings.

Binyu Yang

Mr. Binyu Yang is a PhD student in Translation Studies at The Chinese University of Hong Kong, Shenzhen. He is interested in corpus-based interpreting/translation studies, critical discourse analysis, and the role of interpreters. He holds a MA degree in conference interpreting from the University of Bath and used to be a government in-house interpreter and freelance conference interpreter for many years. He also taught university-level courses at multiple universities in China.

Ye Lei

Ye Lei is a PhD student in interpreting studies at Institute of Corpus Studies and Applications, Shanghai International Studies University, Shanghai, China. He is the author of the of Wordless, an integrated corpus tool with multilingual support for the study of language, literature, and translation. His research interests include corpus-based interpreting studies, corpus-based translation studies, and corpus tool development.

Frances Yung

Frances Yung is a Post-Doc at Saarland University. She has a background in human and machine translation, computational linguistics, and psycholinguistics. Her current research focuses on the multilingual annotation of discourse relations and automatic discourse parsing.

Virginia Zorzi

Virginia Zorzi is a research fellow in English language and linguistics at the University of Bergamo. She received a Ph.D. in Linguistic, Philological and Literary Sciences from the University of Padua. Her

research can be situated within the fields of applied linguistics, discourse analysis and translation studies, and her work chiefly draws on methods related to Corpus Linguistics. Her work concerns discourses on migration in a cross-linguistic and translation perspective. She is also interested in how the public communication of science and technology affects their role and development within society, with a special focus on technoscientific controversies and anti- or pseudoscientific stances, which partly intertwine with conspiracist and/or populist communication.

Jūratė Žukauskaitė

Jūratė Žukauskaitė is a PhD Research Fellow in Translation Studies at the University of Agder, where she explores source language interference (SLI) during simultaneous interpreting from a cognitive linguistic and psycholinguistic perspective, with a particular focus on the relation between lexical SLI and entrenchment.

In 2015, Jūratė graduated from Vilnius University with an MA in conference interpreting and worked as an interpreter until 2022. Her MA thesis investigated note-taking strategies during consecutive interpreting. During her first MA at Humboldt University Berlin, Jūratė specialized in psycholinguistics and examined the relationship between case marking and subject predicate agreement errors in language production.

Edyta Źrałka

Edyta Źrałka, Assistant Professor at the University of Silesia, Katowice

Her main interests are comparative linguistics, specialized languages (mainly legal and political language), and translation studies in both theoretical and practical terms (from the perspective of a practising sworn translator of English and a translation teacher). Manipulation, translation evaluation, metrics in quality assessment of MT and ethics in translation are the topics of her recent articles and a monograph. She has also published articles on localization in translation and the influence of linguistic and musical hearing on learning pronunciation in a foreign language, based on her professional experience in the field of musicology. Her publications include the monograph, journal articles, chapters in edited volumes, and dictionary entries concerning American literary history.

Index

Table of Contents