ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DEPARTMENT OF
INTERPRETING AND TRANSLATION

UCCTS 2021

Using Corpora in Contrastive and Translation Studies

Bertinoro (Italy), 9 - 11 September 2021

Book of Abstracts

Book of Abstracts

Using Corpora in Contrastive and Translation Studies Conference (6th edition)

Sara Castagnoli, Silvia Bernardini, Adriano Ferraresi, Maja Miličević Petrović (eds)

Bertinoro (Italy), 9-11 September 2021

Conference convenors

Silvia Bernardini (Università di Bologna)
Adriano Ferraresi (Università di Bologna)

Organising committee

Ksenia Balakina (Università di Bologna)
Sara Castagnoli (Università di Macerata)
Ester Dolei (Università di Bologna)
Anabela Cristina Costa da Silva Ferreira (Università di Bologna)
Maja Miličević Petrović (Università di Bologna)
Beatrice Ragazzini (Università di Bologna
Natalia Rodríguez Blanco (Università di Bologna)
Mariachiara Russo (Università di Bologna)

Scientific committee

Svetlana Aloushkova (UC Louvain / Université Saint-Louis - Bruxelles)
Alberto Barrón-Cedeño (Università di Bologna)
Łucja Biel (University of Warsaw)
Mario Bisiada (Universitat Pompeu Fabra)
Bert Cappelle (Université de Lille 3)
Sara Castagnoli (Università di Macerata)
Gloria Corpas Pastor (University of Malaga)
Gert De Sutter (Ghent University)
Bart Defrancq (Ghent University)
Isabelle Delaere (KU Leuven)
Ilse Depraetere (Université de Lille 3)
Pamela Faber (University of Granada)
Ana Frankenberg-Garcia (University of Surrey)
Federico Gaspari (University for Foreigners "Dante Alighieri" of Reggio Calabria)
Gaëtanelle Gilquin (UC Louvain)
Sylviane Granger (UC Louvain)
Silvia Hansen-Schirra (Johannes Gutenberg University of Mainz)
Hilde Hasselgård (University of Oslo)
Juliane House (University of Hamburg)
Ilmari Ivaska (University of Turku)
Marta Kajzer-Wietrzny (Adam Mickiewicz University)
Dorothy Kenny (Dublin City University)
Haidee Kotze (Utrecht University)
Natalie Kübler (Université Paris-Diderot)
Kerstin Kunz (University of Heidelberg)
Ekaterina Lapshinova-Koltunski (Saarland University)
Sara Laviosa (University of Bari Aldo Moro)
Marie-Aude Lefer (UC Louvain)
Agnieszka Lenko-Szymanska (University of Warsaw)
Natalia Levshina (Max Planck Institute for Psycholinguistics, Nijmegen)
Defeng Li (University of Macau)
Rudy Loock (Université de Lille 3)
Sandra Halverson (University of Agder)
Juana Isabel Marín Arrese (Universidad Complutense de Madrid)
Josep Marco Borrillo (Universitat Jaume I)

Lorenzo Mastropierro (University of Nottingham)
Anna Mauranen (University of Helsinki)
Adriana Mezeg (University of Ljubljana)
Christine Michaux (Université de Mons)
Tamara Mikolič Južnič (University of Ljubljana)
Maja Miličević Petrović (University of Bologna)
Laura Mori (University of International Studies - Rome)
Ricardo Muñoz Martín (Università di Bologna)
Stella Neumann (RWTH Aachen University)
Raluca Nita (Université de Poitiers)
Signe Oksefjell Ebeling (University of Oslo)
Magali Paquot (UC Louvain)
Giuseppe Palumbo (Università di Trieste)
Koen Plevoets (Ghent University)
Rosa Rabadán (University of León)
Mariachiara Russo (Università di Bologna)
Raf Salkie (University of Brighton)
Erich Steiner (Universität des Saarlandes)
Frieda Steurs (KU Leuven)
Elke Teich (Universität des Saarlandes)
Sonia Vandepitte (Ghent University)
Gudrun Vanderbauwhede (Université de Mons)
Åke Viberg (Uppsala University)
Federico Zanettin (Università di Perugia)
Sandrine Zufferey (Université de Berne)

Table of contents

KEYNOTE PRESENTATIONS

Breaking new ground in contrastive and translation studies: Learner translation corpora to the fore

Marie-Aude Lefer
Université catholique de Louvain
marie-aude.lefer@uclouvain.be

Corpus-based contrastive and translation studies have considerably expanded their reach in recent years, with noteworthy advancement on many fronts – theoretical, methodological and descriptive – and increased interdisciplinarity (see e.g. De Sutter & Lefer 2020, Enghels et al. 2020, Granger & Lefer 2020a, Kotze forthcoming, Neumann et al. forthcoming, Vandevoorde et al. 2020). As confirmed by recent surveys of corpus-based contrastive linguistics (Hasselgård 2020) and translation studies (Granger & Lefer forthcoming), one thing that contrastive and translation studies have in common is their strong reliance on parallel corpora, i.e. corpora containing source texts in a given language, aligned with their translations in another language. These parallel corpora are typically made up of *professional* (or expert) translations into the translators' *native language*, with relatively few exceptions (Lefer 2020). In this talk, I aim to show how corpus-based contrastive and translation studies can break new ground by making use of *learner translation corpora* (LTC). LTC are a subtype of learner corpora (collections of texts produced by learners of a language; Gilquin 2020). More specifically, they are parallel corpora containing novice translations produced, for example, by translation students or foreign language learners. Being instances of both parallel corpora and learner corpora, LTC are situated at the interface of translation studies, contrastive linguistics and learner corpus research (Granger & Lefer 2020b).

Calls for combined analysis of bilingual comparable or parallel data and learner data first emanated from learner corpus research, with Granger's (1996) *Integrated Contrastive Model* (see Gilquin 2000/2001, Granger 2015, 2018 and Hasselgård & Ebeling 2018 for extensions of the model). The basic tenet of the model is that bilingual data can help shed light on transfer in learner language. The potential that lies in such combinations of corpus data types was also acknowledged quite early on in both contrastive linguistics (Johansson 2007: 313) and translation studies (Chesterman 2007: 63). Even though the proposed approaches have so far been taken up only timidly in empirical investigations (e.g. Altenberg 2002, Behrens 2006, Ebeling & Hasselgård 2021, Gilquin 2008, Vanderbauwhede 2012), it is expected that they will soon be thriving, for instance under the impetus of the constrained language framework put forward by Kotze (2020, forthcoming) (see e.g. De Sutter & Lefer 2020, Ivaska et al. forthcoming). LTC represent yet another way of integrating learner data into contrastive and translation studies. To date, however, their use in empirical research has remained relatively marginal.

The first LTC emerged two decades ago (e.g. PELCRA - Uzar & Walinski 2001, STA - Bowker & Bennison 2003) and were soon followed by similar initiatives (e.g. MISTiC - Castagnoli 2009, MeLLANGE - Castagnoli et al. 2011, NEST - Graedler 2013, UPF - Espunya 2014, RusLTC - Kutuzov & Kunilovskaya 2014, KOPTE - Wurm 2016, CELTraC - Fictumova et al. 2017). In this talk, I will sketch out the main features of the LTC compiled to date (language pairs, translation directionality, registers, etc.) and survey existing LTC-based research, in terms of research objectives, topics, corpus methodology and key findings. I will show that the main foci so far have been on computer-aided translation error analysis and translation quality evaluation (e.g. De Sutter et al. 2017, Espunya 2014, Kübler et al. 2018, Vela et al. 2014), mostly with a view to informing translation pedagogy and devising corpus-informed teaching materials. Alongside this core applied-research strand, new types of LTC-based empirical investigation have started to emerge in recent years, especially as regards the study of translation features, which had hitherto been approached mainly through the lens of expert translation (see e.g. Castagnoli 2016, Kunilovskaya et al. 2018, Lapshinova-Koltunski forthcoming, Loock 2020, Redelinghuys & Kruger 2015).

The next part of my talk will be devoted to the *Multilingual Student Translation* (MUST) project, an international LTC collection initiative which brings together more than 40 partner teams worldwide (Granger & Lefer 2020b). Among its many strengths, I will describe its shared source-text database, its rich standardized metadata related to the source texts, translation tasks and learners, and the Translation-oriented Annotation System (TAS) that is currently being developed collaboratively within the MUST network. I will also present some MUST-based studies to illustrate the research potentialities of the corpus (e.g. Penha-Marion et al. forthcoming on translation directionality).

The talk will end with a few forward-looking remarks. In particular, I will discuss the various ways in which LTC can be combined with other corpus and observational data types to shed new light on cross-linguistic contrasts and translation. I will also argue that LTC data can help foster theoretical development, especially as regards the further elaboration of key constructs such as translation competence acquisition (e.g. through longitudinal studies; see Wurm 2020), translation expertise (e.g. Kajzer-Wietrzny 2020), default translation (Halverson 2019) and translation variation (e.g. Castagnoli 2020). Finally, I will sketch out the promising new opportunities for LTC-based research to remain relevant in today's world, such as the collection of LTC devoted to forms of interlingual mediation other than written translation (e.g. post-editing, subtitling, video game and web localization).

References

Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 37-54.

Behrens, B. (2006). Language-based processing in advanced L2 production and translation: An exploratory study. In H. Byrnes, H. Weger-Guntharp & K. Sprang (eds.) *Educating for Advanced Foreign Language Capacities*. Georgetown University Press, 74-86.

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

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*. Corpora and Language in Use Proceedings 6. Louvain-la-Neuve: Presses universitaires de Louvain, 25-44.

Castagnoli, S. (2016). Investigating trainee translators' contrastive pragmalinguistic competence: a corpus-based analysis of interclausal linkage in learner translations. *The Interpreter and Translator Trainer* 10(3), 343-363.

Castagnoli, S. (2009). *Regularities and variations in learner translations: A corpus-based study of conjunctive explicitation*. PhD thesis, Pisa University.

Castagnoli, S., Ciobanu, D., Kübler, N., Kunz, K., & Volanschi, A. (2011). Designing a Learner Translator Corpus for Training Purposes. In N. Kübler (ed.) *Corpora, Language, Teaching, and Resources: From Theory to Practice*. Bern: Peter Lang, 221-248.

Chesterman, A. (2007). Similarity analysis and the translation profile. *Belgian Journal of Linguistics* 21, 53-66.

De Sutter, G., Cappelle, B., De Clercq, O., Loock, R., & Plevoets, K. (2017). Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 16, 25-39.

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.

Ebeling, S. & Hasselgård, H. (2021). The functions of n-grams in bilingual and learner corpora: An integrated contrastive approach. In S. Granger (ed.) *Perspectives on the L2 Phrasicon: The View from Learner Corpora*. Multilingual Matters.

Enghels, R., Defrancq, B., & Jansegers, M. (eds.) (2020). *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Berlin: De Gruyter Mouton.

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation* 48, 33-43.

Fictumova, J., Obrusnik, A., & Stepankova, K. (2017). Teaching specialized translation error-tagged translation learner corpora. *Sendebar* 28, 209-241.

Gilquin, G. (2020). Learner corpora. In M. Paquot & S. Th. Gries (eds.) *A Practical Handbook of Corpus Linguistics*. Springer, 283-303.

Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In G. Gilquin, S. Papp & B. Díez-Bedmar (eds.) *Linking Contrastive and Learner Corpus Research*. Amsterdam & New York: Rodopi, 3-33.

Gilquin, G. (2000/2001) The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast* 3(1), 95-124.
Graedler, A.-L. (2013). NEST—A corpus in the brooding box. *Studies in Variation, Contacts and Change in English*, 13.
Granger, S. (2018). Tracking the third code: A cross-linguistic corpus-driven approach to metadiscursive markers. In A. Cermakova & M. Mahlberg (eds.) *The Corpus Linguistics Discourse*. Amsterdam & Philadelphia: John Benjamins, 185-204.
Granger, S. (2015). Contrastive Interlanguage Analysis. A reappraisal. *International Journal of Learner Corpus Research* 1(1), 7-24.
Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (eds.) *Languages in Contrast: Text-based Cross-Linguistic Studies. Lund Studies in English*. Lund: Lund University Press, 37-51.
Granger, S. & Lefer, M.-A. (forthcoming). 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.
Granger, S. & Lefer, M.-A. (eds.) (2020a). *The Complementary Contribution of Comparable and Parallel Corpora to Crosslinguistic Studies*. Special issue of *Languages in Contrast*, 20(2).
Granger, S. & Lefer, M.-A. (2020b). The *Multilingual Student Translation* corpus: a resource for translation teaching and research. *Language Resources and Evaluation* 54, 1183-1199.
Halverson, S. L. (2019). 'Default' translation. A construct for cognitive translation and interpreting studies. *Translation, Cognition & Behavior* 2(2), 187-210.
Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments and directions. In S. Granger & M.-A. Lefer (eds.) *The Complementary Contribution of Comparable and Parallel Corpora to Crosslinguistic Studies*. Special issue of *Languages in Contrast* 20(2), 184-208.
Hasselgård, H. & Ebeling, S. (2018). At the interface between Contrastive Analysis and Learner Corpus Research: A parallel contrastive approach. *Nordic Journal of English Studies* 17(2),182-214.
Ivaska, I., Ferraresi, A., & Bernardini, S. (forthcoming). Syntactic properties of constrained English: A corpus-driven approach. In S. Granger & M.-A. Lefer (eds.) *Extending the Scope of Corpus-based Translation Studies*. Bloomsbury.
Johansson, S. (2007). *Seeing through Multilingual Corpora On the use of corpora in contrastive studies*. Amsterdam and Philadelphia: John Benjamins.
Kajzer-Wietrzny, M. (2019). Linking words in inter- and intralingual translation - Combining corpus linguistics and key-logging data. In L. Vandevoorde, J. Daems & B. Defrancq (eds.) *New Empirical Perspectives on Translation and Interpreting*. Abingdon: Routledge, 114-138.
Kotze, H. (forthcoming). Translation as constrained communication: Principles, concepts and methods. In S. Granger & M.-A. Lefer (eds.) *Extending the Scope of Corpus-based Translation Studies*. Bloomsbury.
Kotze, H. (2020). Converging what and how to find out why: An outlook on empirical translation studies. In L. Vandevoorde, J. Daems & B. Defrancq (eds.) *New Empirical Perspectives on Translation and Interpreting*. Abingdon: Routledge, 333-370.
Kübler, N., Mestivier-Volanschi, A., & Pecman, M. (2018). Teaching specialised translation through corpus linguistics: quality assessment and methodology evaluation by experimental approach. *Meta* 63(3), 806-824.
Kunilovskaya, M., Morgoun, N., & Pariy, A. (2018). Learner vs. professional translations into Russian: Lexical profiles. *Translation and Interpreting* 10(1), 33-52.
Kutuzov, A., & Kunilovskaya, M. (2014). Russian learner translator corpus: design, research potential and applications. In P. Sojka, A. Horak, I. Kopecek & K. Palak (eds.) *Text, Speech and Dialogue. Lecture Notes in Computer Science*. Berlin: Springer, 315-323.
Lapshinova-Koltunski, E. (forthcoming). Detecting normalisation and shining-through in novice and professional translations. In S. Granger & M.-A. Lefer (eds.) *Extending the Scope of Corpus-based Translation Studies*. Bloomsbury.
Lefer, M.-A. (2020). Parallel corpora. In M. Paquot & S. Th. Gries (eds.) *A Practical Handbook of Corpus Linguistics*. Springer, 257-282.
Loock, R. (2020). It's non-canonical word order that you should use! A corpus approach to avoiding standardized word order in translated French. In S. Granger & M.-A. Lefer (eds.) *Translating and Comparing Languages: Corpus-based Insights*. Corpora and Language in Use Proceedings 6. Louvain-la-Neuve: Presses universitaires de Louvain, 69-85.
Neumann, S., Freiwald, J., & Heilmann, A. (forthcoming). On the use of multiple methods in empirical translation studies: A combined corpus and experimental analysis of subject identifiability in English and German. In S. Granger & M.-A. Lefer (eds.) *Extending the Scope of Corpus-based Translation Studies*. Bloomsbury.
Penha-Marion, L. A. de S., Gilquin, G., & Lefer, M.-A. (forthcoming). Lexico-syntactic simplification in French><English student translation: The effect of translation directionality on bilingual language production. In H. Kotze & B. Van Rooy (eds.) *Constraints on Language Variation and Change in Complex Multilingual Contact Settings*. John Benjamins.
Redelinghuys, K. & Kruger, H. (2015). Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics* 20(3): 293-325.
Uzar, R. & Walinski, J. (2001). Analysing the fluency of translators. *International Journal of Corpus Linguistics* 6, 155-166.
Vanderbauwhede, G. (2012). The Integrated Contrastive Model evaluated: the French and Dutch demonstrative determiner in L1 and L2. *International Journal of Applied Linguistics* 22(3), 392-413.
Vandevoorde, L., Daems, J., & Defrancq, B. (eds.) (2020). *New Empirical Perspectives on Translation and Interpreting*. Abingdon: Routledge.

Vela, M., Schumann, A.-K., & Wurm, A. (2014). Beyond linguistic equivalence. An empirical study of translation evaluation in a translation learner corpus. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, 47-56.

Wurm, A. (2020). Translation quality in an error-annotated translation learner corpus. In S. Granger & M.-A. Lefer (eds.) *Translating and Comparing Languages: Corpus-based Insights*. Corpora and Language in Use Proceedings 6. Louvain-la-Neuve: Presses universitaires de Louvain, 141-162.

Wurm, A. (2016). Presentation of the KOPTE Corpus and Research Project. https://www.academia.edu/24012369/Presentation_of_the_KOPTE_Corpus_and_Research_Project.

Corpora and cross-linguistic comparison: opportunities and challenges

Natalia Levshina
Max Planck Institute for Psycholinguistics, Nijmegen
Natalia.Levshina@mpi.nl

Over the last few years, the number of corpora that can be used for language comparison has dramatically increased. In my talk I will chart this new and changing territory, providing a few landmarks, warning signs and safe paths. Although no corpus at present can replace the traditional type of typological data based on language description in reference grammars, corpora can help with diverse tasks, being particularly well suited for investigating probabilistic and gradient properties of languages and for discovering and interpreting cross-linguistic generalizations based on processing and communicative mechanisms. At the same time, the use of corpora for typological purposes has not only advantages and opportunities, but also numerous challenges. I will also present several empirical case studies, including one addressing pertinent problems, namely, the role of text types in language comparison and the problem of using the word as a cross-linguistic comparative concept.

Translation as specialised language use. A probabilistic view of linguistic characteristics of translations

Stella Neumann
RWTH Aachen University
stella.neumann@ifaar.rwth-aachen.de

Translation is simply language use, albeit specialised language use. If it wasn't, translations would be rejected as incomprehensible in the target language (TL). In many cases, the translation will be covert (House 1997) intended to work in the same way as a non-translated text. Against this background, we would expect translations to be indistinguishable from texts produced in the TL without an anterior text (Halverson 2013) in the source language. Moreover, the (non-translated) texts produced in the TL will also include other types of texts produced under conditions of language contact, i.e. involving a second language in addition to the TL. It therefore appears implausible to expect any linguistic peculiarities in (covert) translations. Yet, many computational studies have reported high accuracies in classifying translated texts based on their linguistic features (for an early example, see Baroni and Bernardini 2006). So, there must be something that makes translations easy to spot for the computer. In this paper, I will review the corpus-based approach to translation adopting a probabilistic view of language. Translation is a specialised form of language use because of its link to an anterior text: information linguistically encoded in the anterior text has to be re-coded in a different language system (Hansen-Schirra and Steiner 2012; similarly Halverson 2013). Hansen-Schirra and Steiner claim that translation is arguably the only type of text production which is linked to such a previous encoding – at least when excluding "[a]ny weaker form of multilingual text production in the sense of producing target context-adapted re-creation" (2012, 261). If their claim is right, precisely this link may be the reason for the specific distribution of linguistic features to which the computational classification task responds. Pressure on the translator to re-encode the anterior text's meaning and wording might lead to observable linguistic differences from non-translated TL texts, while attempting to cover up the fact that someone else has expressed this meaning previously in a different language might result in differences from the encoding of the anterior text and, more generally, the source language.

The by now widely accepted view of language as a dynamic, open, that is, probabilistic system (e.g. Halliday 1991, Beckner et al. 2009, see also Toury 2004) offers an explanation for the uneven, yet systematic distribution of linguistic features across texts produced under different conditions. This view implies a paradigmatic conceptualisation of language according to which language users have a range of more or less likely linguistic options at their disposal for expressing a certain meaning. Since translation is language use it is subject to the same types of systematic variation in probabilities that also apply to non-translated texts. The translator as language user consequently also chooses between various options even if the default translation is blocked, resulting in translation shifts, thus explaining variation between translations. The choice is systematically influenced by factors such as situational context, but also social and cognitive factors. More specifically, the findings of the above mentioned computational studies suggest that the influence of various factors applies differently in translations as compared to non-translated texts. Consequently, as suggested by De Sutter and Lefer (2020), the task of empirical translation studies is to ascertain the influence of the multiple factors that condition the outcome of the translation process. From a probabilistic point of view, the complex influence of factors affecting translational language use means that the effect is discernible, but weak. Crucially, this also means that translations will be gradually, not categorically different from other types of text. This has consequences not only for the corpus methodology in translation studies, thus aligning with De Sutter and Lefer's (2020) call for methodological innovation, but also for the way we interpret the results of corpus studies.

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), 259-74.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning* 59(s1), 1-26.

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.

Halliday, M. A. K. (1991). Towards Probabilistic Interpretations. In E. Ventola (ed.) *Functional and Systemic Linguistics. Approaches and Uses*. Berlin, New York: Mouton de Gruyter, 39–61.

Halverson, S.L. (2013). Implications of Cognitive Linguistics for Translation Studies. In A. Rojo & I. Ibarretxe-Antunano (eds.) *Cognitive Linguistics and Translation. Advances in Some Theoretical Models and Applications*. Berlin, Boston: de Gruyter, 33-74.

Hansen-Schirra, S. & Steiner, E. (2012). Towards a Typology of Translation Properties. In S. Hansen-Schirra, S. Neumann & E. Steiner, *Cross-Linguistic Corpora for the Study of Translations*. Berlin: De Gruyter, 255-279.

House, J. (1997). *Translation Quality Assessment. A Model Revisited*. Tübingen: Gunter Narr Verlag.

Toury, G. (2004). Probabilistic Explanations in Translation Studies. Welcome as They Are, Would They Qualify as Universals? In A. Mauranen & P. Kujamäki (eds.) *Translation Universals. Do They Exist?* Amsterdam/Philadelphia: John Benjamins Publishing Company, 15–32.

FULL PAPERS

Interpreting polemical sequences: A corpus study based on debates of the European Parliament about the situation in Poland

Magdalena Bartłomiejczyk
University of Silesia in Katowice
ogien@o2.pl

The European Union relies heavily on translation and interpreting in its everyday functioning. During plenary debates of the European Parliament (EP), the voices of Members sound in the original version in the language selected by the speaker and in 23 interpretations into all the other EU languages. The voice of each speaker, therefore, inevitably becomes "filtered" by 23 other individuals. Over recent years, numerous empirical studies have generated many insights into the multilingual functioning of the EP. Some of the studies (e.g. Beaton-Thome 2013, Bartłomiejczyk 2016, Kučiš and Majhenič 2018) have shown that interpreters might introduce more serious shifts than assumed by the conduit model of interpreting and that the filtering effect may sometimes considerably modify the speaker's illocution.

This study employs discourse analysis for a qualitative, pragmatically-oriented exploration of polemical sequences extracted from a new corpus that I am currently compiling. It includes the plenary debates devoted specifically to the developments in Poland since 2015, i.e. the time when the Law and Justice Party came to power and began introducing very controversial reforms, primarily of the judicial system. Overall, the corpus now contains 9 debates and the relevant explanations of vote (contributions in English and in Polish and their interpretations into the other language).

Simultaneous interpreting is mainly associated with interlingual transfer of monologic discourse rather than of interactions among the participants. Interpreting plenary debates of the EP also fits into this pattern. As rightly argued by Kent, "Although described as 'debate', the speeches given by Members during plenaries are mainly directed to consumption by home country audiences via the internet, television and radio rather than as engagement with colleagues who are in the same room" (2009: 57). Marzocchi (1998) also notes lack of spontaneity and little interaction among speakers participating in the plenary, and points out that real discussions are conducted in other types of meetings with scarce public exposure. Consequently, the material under analysis here is an exception to the general rule of monologism and may constitute input to which the interpreters are not necessarily accustomed.

Simultaneous interpreting of interactions (particularly if the original contributions are in two or more different languages) bears some similarity to liaison interpreting as a triadic exchange. At the same time, the interpreter's role as a mediator and coordinator is limited due to his/her spatial distancing from the participants and no possibility of addressing them directly for the sake of managing the interaction (in contrast to liaison interpreting, see e.g. Wadensjö 1998). In particular, I examine features such as intertextual references to contributions of other Members, personal references (including deixis), terms of address and face-threatening acts targeted at interlocutors. These are studied to detect possible shifts introduced by interpreters.

For example, the debate on the rule of law and democracy in Poland held on 15 November 2017 contains a verbal scuffle between two Polish MEPs: Janusz Lewandowski and Ryszard Legutko (who both speak Polish). This interaction is then joined by Guy Verhofstadt (speaking English and presumably listening to the English interpretations of the Polish contributions), who refers to both the Polish MEPs in his contribution:

Mr Legutko, why you don't stay for the debate? Why you are leaving? No, because I have to say something to you. I have to say something to you. Your attack to Mr Lewandowski I find it outrageous. If there is one sensible… If there is one sensible, reasonable, colleague of us, sometimes even a little bit boring, then it is Mr Lewandowski. And to say that he has lose his senses, well, I think it's the Polish Government that has lost his senses and not Mr Lewandowski.

There are some interesting shifts in the interpretation of this fragment as well as subsequent parts of the speech. The interpreter seems to have missed the direct address to Legutko at the beginning (possibly due to the long time lag resulting from waiting to take over from her boothmate) and she apparently construes the pronoun *you* as plural, which is justified by the situational context (the leaving MEP is accompanied by other persons). In the next sentence, she uses the 3rd person singular form *pan poseł* 'Mr MEP', which introduces some ambiguity (it may be understood either as a direct, polite question to Legutko or as a rhetorical question to the audience). The term of address fully conforms to the rules of Polish grammar, and, moreover, Polish politeness that prescribes mentioning positions rather than names (the same pragmatic adjustment is applied consistently also to the reference to Lewandowski). However, omitting the name makes it difficult to specify the addressee, and probably this is why the interpreter adds the explanation *Zwracam się do pana posła Legutko* 'I'm addressing Mr MEP Legutko'. On the whole, the Polish version of this part is more indirect and less insistent (in other words, more polite) than the original.

As for the reference to Lewandowski, a similar trend towards indirectness is visible when the interpreter renders *colleague of us* with the less personal *kolega* without any possessive pronouns. The reported accusation towards Lewandowski is transferred twice *postradał zdrowy rozsądek i postradał zmysły* 'he lost common sense and he lost his senses', which looks like a search for a more accurate translation rather than strengthening of criticism. This accusation is probably the most interesting element as such, because in the original Polish contribution to which Verhofstadt is referring Legutko did not question Lewandowski's sanity, he accused him of being utterly immoral and telling lies. This particular face threat was introduced by the English interpreter, who added *he seems to have lost control of his senses.* Consequently, we see very clearly how the interpreter's shift influences the course of the debate, as Verhofstadt is obviously defending Lewandowski against what the interpreter said and not what Legutko said. I would like to present several similar examples in my paper in Bologna.

References:

Bartłomiejczyk, M. (2016). *Face threats in interpreting. A pragmatic study of plenary debates in the European Parliament.* Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Beaton-Thome, M. (2013). What's in a word? Your 'enemy combatant' is my 'refugee'. The role of simultaneous interpreters in negotiating the lexis of Guantánamo in the European Parliament. *Journal of Language and Politics* 12 (3), 378–399.

Kent, S. J. (2009). A discourse of danger and loss. Interpreters on interpreting for the European Parliament. In: L. Stern, U. Ozolins and S.B. Hale (Eds.) *The Critical Link 5: Quality in Interpreting: A Shared Responsibility.* Amsterdam and Philadelphia: Benjamins, 55–70.

Kučiš, V. and Majhenič, S. (2018). Cultural and stress-related manifestations of political controversial language in the European Parliament from the view of interpreters. *Babel* 64 (1), 33–62.

Marzocchi, C. (1998). The case for an institution-specific component in interpreting research. *The Interpreters' Newsletter* 8, 51–74.

Wadensjö, C. (1998). *Interpreting as interaction.* London and New York: Longman.

Terminological collocations in trainee legal translations: a learner-corpus study of L2 company law translations

Łucja Biel, Agnieszka Leńko-Szymańska
University of Warsaw
l.biel@uw.edu.pl, a.lenko@uw.edu.pl

One of the fundamental challenges of legal translation is the system-bound nature of legal terms (cf. Šarčević, 1997), that is their conceptual embedding in a given legal system, and the resulting incongruity of legal terms between source and target systems. While most attention is devoted to terms, it needs to be remembered that the system-bound nature is shared by legal phraseology, in particular collocations of terms which embed terms in text (term-embedding collocations or terminological collocations). Terminological collocations are important building blocks in legal discourse, mapping elements of conceptual frames (Meyer & Mackintosh, 1994, p. 346) and legal rules. Due to the high formulaicity and 'petrification' of legal language (Crystal and Davy,1969, p. 194); Mattila, 2006, p. 233), legal collocations are more fixed and restricted, showing lower variation and synonymy than general-language and other LSP collocations (Biel, 2014).

Researchers in second language acquisition generally agree that L2 learners are slow at acquiring productive knowledge of L2 collocations (e.g., Durrant & Schmitt, 2009; Nekrasova, 2009). In fact, lack of idiomaticity is one of the most prominent indicators of non-nativeness and persists even at advanced levels of proficiency. Two groups of factors have been proposed to explain this phenomenon. The first one relates to the very nature of phraseology: low frequency of individual word combinations and their lack of semantic and perceptual saliency. The latter group is linked to the characteristics of L2 learners, who tend to focus on individual words rather than multi-word chunks in their learning and who lack awareness of the importance of idiomaticity in language (Boers et al., 2014).

However, the SLA literature to date has devoted little attention to the acquisition of specialized phraseology – terminological collocations – by a special kind of L2 learners – translation trainees. The issue worth pursuing in this context is whether translation students' awareness of multiword-nature of legal terminology and the fixedness and restrictiveness of its phraseology has a facilitating effect on the acquisition of L2 word combinations and on the adequacy of their choices of terminological collocations.

The aim of this paper is to analyse empirically, using corpus methodology, the rendering of terminological collocations in the Polish-English L2 translations of *statut*, one of the key foundational documents of public limited companies. We are interested in exploring if, and to what extent, students' solutions differ from those of expert translators and L1 professionals, in particular as regards the range and variation of collocational choices. Other analysed phenomena will include domestication/foreignization and adequacy/acceptability of terminological collocations in translated texts, as well as the types of errors.

The data will be drawn from the following sources: (1) a focus corpus: a Polish-English section of the MUST learner corpus with translations rendered by MA Translation Programme students at the University of Warsaw; (2) a comparable parallel corpus of expert translations WIG-20 containing English translations of Articles of Associations of top 20 Polish listed companies; (3) two comparable corpora of nontranslated Articles of Association of top 20 UK and US listed companies as a benchmark.

The study will focus on two kinds of verbal collocations: *subject + verb* and *verb + object,* where both subject and object are represented by one- and multi-word legal terms. The first step in the analysis will involve extracting terminological noun phrases from the learner and the expert translation corpus as well

as the benchmark corpora. Next, verbal collocation of these terms in the four corpora will be analysed and juxtaposed both qualitatively and quantitively. Retrieved collocations will be classified into four categories: (1) adequate terminological collocations; (2) acceptable but rare collocations; (3) collocations from informal or non-specialised registers; and (4) calqued collocations.

The results of the study may provide an insight into the process of learning specialized language by translation trainees. It may also provide useful implications for translation training.

References

Biel, Ł. (2014). *Lost in the Eurofog. The Textual Fit of Translated Law*. Frankfurt am Main: Peter Lang.

Boers, F., Lindstromberg, S., & Eyckmans, J. (2014). Some explanations for the slow acquisition of L2 collocations. *Vigo International Journal of Applied Linguistics* 11, 41–62.

Crystal, D., & Davy, D. (1969). *Investigating English Style*. London: Longman.

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

Mattila, H. E. S. (2006). *Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Franca.* 2nd Edition. New York: Routledge.

Meyer, I., & Mackintosh, K. (1994). Phraseme Analysis and Concept Analysis: Exploring a Symbiotic Relationship in the Specialized Lexicon. In W.E.A. Martin (Ed.), *Euralex 1994. Proceedings*. Amsterdam: International Congress on Lexicography, 339–348.

Nekrasova, T. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning* 59, 647–686.

Newmark, P. (1988). *A Textbook of Translation*. London: Prentice Hall.

Šarčević, S. (1997). *New Approach to Legal Translation*. The Hague: Kluwer Law InternationalBaker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research. *Target* 7(2), 223-243.

The Polish Eurolect across genres: an internal and external variation in EU translation

Łucja Biel, Katarzyna Wasilewska, Dariusz Koźbiał
University of Warsaw
l.biel@uw.edu.pl, k.e.wasilewska@gmail.com, d.kozbial@uw.edu.pl

EU (European Union) discourses involve an unprecedented degree of mediation by translators and filtering through 24 official languages, which results in their hybridity and "the extreme visibility of the 'translatedness' of the texts" (Koskinen 2000, 61). Combined with a complex array of institutional, political, procedural and supranational factors, the hybridity results in an emergence of "Europeanized" parallel varieties of national legal and administrative languages, known as Eurolects. Eurolects have developed a distinct supranational terminology, as well as stylistic and grammatical features, which depart from certain conventions of national languages, partly due to the Eurolects' extreme mutual filtering through other EU languages (cf. Biel, 2014). The feeling that Eurolects are 'different' is evidenced by their common discursive construction and stigmatization as Other: Eurospeak, Eurojargon, Eurolanguage, Euro-rhetoric, Euro-Legalese, Union legalese, EUese, Eurofog, Eurish (cf. Goffin (1994, 636); Biel (2014, 76)). With the advent of corpus methods, it has recently become possible to explore this difference empirically on a large scale. First corpus studies into the peculiarities of EU language were conducted in the 2000s to investigate the selected lexico-grammatical patterns of EU English (e.g. Foley (2001), (2002); Caliendo (2004), Caliendo, Martino, and Venuti (2005); Caliendo (2007), Mazzi (2007), Trebits (2009)). With time, they evolved to embrace other languages (e.g. the Eurolect Observatory Project which covers 11 languages, (Mori 2018b)) and shifting from microstructures to holistic macro-level studies investigating Eurolects as a whole from the perspective of Translation Studies (Biel 2014); and (2) sociolinguistics and contact linguistics (Mori 2018a) (cf. Biel (*forthcoming*) for an overview).

This paper will report on a 4-year project, The Polish Eurolect Project, synthesising its findings and discussing implications for further studies into the nature of Eurolects. The main objectives of the project was to investigate the nature of the Polish Eurolect across four administrative genres to understand the processes and factors behind its formation, and (2) to track its impact on post-accession domestic Polish. Although frequently thought of as a monolith, the Eurolect subsumes a broad range of genres and our analysis covered four legal and semi-legal genres: legislation, judgements, administrative reports and websites for citizens. The first objective was researched through: (1) external variation (textual fit): How does the Eurolect differ from naturally occurring administrative Polish? (2) internal genre-based variation: How does the Eurolect differ internally across four genres (legislation, judgments, reports, official websites for citizens) and diachronically due to the institutionalisation of the translation process (pre-accession versus post-accession translations). The second objective was researched by comparing pre-accession Polish (1999/2000) and post-accession Polish (2015) to measure its "Europeanisation" as a result of the huge inflow of EU translations. In this paper we will focus on demonstrating how the Polish Eurolect differs across genres, that is how a variable of genre affects the nature of translated language and its distance to comparable domestic administrative language.

To answer these research questions we built a large genre-controlled comparable-parallel corpus. The corpus covers four subcorpora of EU legislation, judgments, reports and institutional websites for citizens (both English and Polish language versions) and corresponding comparable corpora of domestic genres (reference corpora). The corpora were designed according to the same sampling frame spanning 5 years, from 2011 to 2015, except for websites which were collected as at 2015/2016. Thus, each genre is represented by three corpora: the main EU Polish corpus, its corresponding EU English corpus and the reference corpus of domestic non-translated texts. In order to control some variables and as a corrective

measure, the corpus of legislation and the corpus of judgments are divided into two components with subgenres (regulations and directives; judgments of the Court of Justice and judgments of the General Court, respectively). To ensure comparability, non-normative preambles and technical annexes were deleted from the EU legislation corpora, which now contains only enacting terms (the normative part). A more detailed corpus description may be found in Biel (2016) and Biel, Koźbiał, and Wasilewska (2019). We used both Wordsmith Tools 7.0 (Scott 2016) and Sketchengine (Kilgarriff et al. 2014).

The study applies mainly comparable corpus methods, starting with the analysis of keywords and word lists to understand how each of our translated EU genre diverge from their national varieties. Keywords are verified through concordances and qualitative manual analyses. To better understand the implications of such divergences, we attempt to do genre profiling to identify key markers for each genre. Those (mainly grammatical and stylistic) features are next compared across genres through multidimensional analysis. The next step was to study selected features on a parallel corpus to understand which source-language features trigger divergences.

Other methods involved the analysis of lexical bundles (cf. Biber and Barbieri 2007), demonstrating a strong correlation between formulaicity and genres, as well as multiple facets of formulaicity (e.g. tokens vs. types). Our findings generally confirm the increased aggregate formulaicity of translations as regards bundle tokens for all EU genres, except for judgments, and the increased variation of bundles (types) for all the genres. Another finding reveals a consistently low overlap of bundles between translations and non-translations. We argue that translations develop their own formulaic profiles which are levelled out compared to EU English corpora and which minimally overlap with formulaic profiles of domestic genres.

References

Biel, Ł. (2014). *Lost in the Eurofog. The Textual Fit of Translated Law*. Frankfurt am Main: Peter Lang.

Biel, Ł. (2016). Mixed corpus design for researching the Eurolect: a genre-based comparable-parallel corpus in the PL EUROLECT project. In E. Gruszczyńska & A. Leńko-Szymańska (eds.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warszawa: Instytut Lingwistyki Stosowanej, 198-208.

Biel, Ł. (2020). Eurolects and EU Legal Translation. In M. Ji & S. Laviosa (Eds.), *The Oxford Handbook of Translation and Social Practices*: Online: Oxford University Press.

Biel, Ł., Koźbiał, D., & Wasilewska, K. (2019). The formulaicity of translations across EU institutional genres: A corpus-driven analysis of lexical bundles in translated and non-translated language. *Translation Spaces* 8(1), 67-92. doi:https://doi.org/10.1075/ts.00013.bie

Caliendo, G. (2004). EU Language in Cross-Boundary Communication. *Textus* 17, 159–178.

Caliendo, G. (2007). Modality and Communicative Interaction in EU Law. In C. N. Candlin & M. Gotti (Eds.), *Intercultural Aspects of Specialized Communication*. Bern: Peter Lang, 241-259.

Caliendo, G., Martino, G. D., & Venuti, M. (2005). Language and Discourse Features of EU Secondary Legislation. In G. Cortese & A. Duszak (eds.), *Identity, Community, Discourse: English in Intercultural Settings*. Bern: Peter Lang, 381-404.

Foley, R. (2001). Going out of style? Shall in EU legal English. *UCREL Technical Papers* 13, 185-195.

Foley, R. (2002). Legislative Language in the EU: The Crucible. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique* 15, 361-374. doi:10.1023/A:1021203529151

Goffin, R. (1994). L'eurolecte : oui, jargon communautaire : non. *Meta* 39(4), 636-642. doi:https://doi.org/10.7202/002930ar

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

Koskinen, K. (2000). Institutional Illusions. Translating in the EU Commission. *The Translator* 6(1), 49-65. doi:10.1080/13556509.2000.10799055

Mazzi, D. (2007). The Construction of Argumentation in Judicial Texts: Combining a Genre and a Corpus Perspective. *Argumentation* 21(1), 21-38. doi:10.1007/s10503-007-9020-8

Mori, L. (2018a). Introduction: The *Eurolect Observatory Project*. In L. Mori (ed.), *Observing Eurolects: Corpus analysis of linguistic variation in EU law*. Amsterdam, Philadelphia: John Benjamins, 1-26.

Mori, L. (Ed.) (2018b). *Observing Eurolects: Corpus analysis of linguistic variation in EU law*. Amsterdam, Philadelphia: John Benjamins.

Scott, M. (2016). *WordSmith Tools version 7*. Stroud: Lexical Analysis Software.

Trebits, A. (2009). Conjunctive cohesion in English language EU documents – A corpus-based analysis and its implications. *English for Specific Purposes* 28(3), 199-210. doi:10.1016/j.esp.2009.04.004

How Surprising are Translators Depending on the Competence and Context?

Yuri Bizzoni, Ekaterina Lapshinova-Koltunski
Saarland University
yuri.bizzoni@uni-saarland.de, e.lapshinova@mx.uni-saarland.de

1 Introduction

The present paper deals with a computational analysis of translationese in professional and student English-to-German translations belonging to different registers. While translationese was extensively discussed in the area of corpus-based translation studies and machine translation (MT), there are relatively few computational studies that focus on the relation between translators' level of expertise and translationese throughout different registers. Our primary focus is on register diversification (Biber et. al., 1998) vs. convergence (Kruger and van Rooy, 2012), reflected in the use of constructions with a higher vs. lower perplexity score.

As translationese is probabilistic in nature (Toury, 2004), we build upon an information-theoretical approach. We use a framework that enables a probabilistic design of language use in the form of a language model. We test translation conformity to source and target language in terms of a neural language model's perplexity over Part of Speech (PoS) sequences. We then compare the results of our perplexity measures with the distribution of different PoS patterns across registers to qualitatively analyze translation divergence in the data. Through this approach, we aim at testing two related hypotheses:

Hypothesis 1

We expect professional translators to be more efficient at reproducing the patterns of their target language. If this is the case, we would expect professional translations to elicit lower perplexity scores from the target language model.

Hypothesis 2

On the other hand, students could converge more on standard patterns: due to their lack of expertise, they might have lower register sensitivity, and thus they could be less bold and more repetitive in their use of grammatical constructions. A higher value of perplexity for a register means a less usual (hence, more perplexing) order of PoS with respect to a reference corpus.

2 Methods

We use a dataset of English-to-German translations produced by both professionals and students from the corpora CroCo (Hansen-Schirra et al., 2012) and VARTRA (Lapshinova-Koltunski, 2013). Professional and student translations have common text sources, and represent, therefore, translation variants of the same texts. Our dataset covers seven registers: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to shareholders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). We also use comparable German non-translated texts covering the same seven registers[1]. All texts in the data were automatically tokenized, lemmatized, and annotated with PoS information based on the Universal Dependency framework (Straka and Strakova, 2017) to ensure the comparability of the results in the source and the target languages.

We model language conventions in terms of PoS sequences through Long Short-Term Memory (LSTM), a recurrent neural network architecture, using monolingual corpora of non-translations in both source and

[1] The texts were also exported from the above mentioned corpus CroCo.

target language as a training set. We then test how students' and professionals' translations conform to linguistic conventions using our models' perplexity scores. A higher value of perplexity for a text means a less usual (hence, more surprising) order of PoS sequences with respect to a reference corpus. We expect perplexity values for the professional translations to be lower than for the student translations (Hypothesis 1). In terms of register diversification in the translated data, the essential idea is that a language model trained on a diverse set of registers[2] will find, on average, a converging translation less perplexing, since it contains grammatical structures typical of what we could call "general language". Convergence will result in the homogeneity of perplexity values across different registers. Here, we expect a higher homogeneity, and hence convergence, for students than professionals (Hypothesis 2).

We also compare our perplexity results with the distributions of PoS n-grams[3] across registers and corpora. If students have an accentuated tendency to converge, they should show less diversity than professionals, which is especially revealing given that both professionals and students are translating the same source text, starting from the same source-structures.

3 Results

Hypothesis 1 The results of the model performance on all the four subcorpora under analysis (including significance test (t- and p-value).) in Table 1 show the English model to be less perplexed by professional translations (11.36) than by non-professional ones (12.51). In this way, professionals seem to be closer to their source texts (interference).

	EO-LM	GO-LM	t-value	p-value
EO	8.88	15.08	-11.6	<0.001
GO	11.12	5.93	23.5	<0.001
ST	12.51	11.12	3.2	0.001
PT	11.36	14.39	-10.1	<0.001

Table 1: Perplexity of the English-trained (EO-LM) and the German-trained models (GO-LM) on EO, GO, ST, and PT.

Student translations elicit a higher perplexity score (12.51), which indicates that they are even more surprising for the English model than the comparable German non-translations and translations by professionals, which indicates over-normalization – exaggerating the target language patterns. The German model's results reveal an opposite tendency: professional translations seem to be more perplexing to the German model than the student ones. Interpreting this result in terms of translationese, such a high level of perplexity, not far from the perplexity reached by English data, could indicate a degree of interference in professionals. This tendency is against our expectations formulated in Hypothesis 1.

Hypothesis 2 The results in Table 2[4] show that almost all registers translated by professionals elicit higher scores than those translated by students. We interpret the lower scores of student translations as a reduced register distinction in favor of a more general language, which confirms our hypothesis that students are more repetitive in the language constructions they use. This can be explained either by the lack of the register-specific knowledge or repetition of specific transfer patterns by students. Because they tend to repeat the same patterns for different registers, students seem less perplexing than professionals. We verify these assumptions in the experiments on pattern diversity.

[2] We trained language models on the texts of the target language corpus that contain all registers.

[3] We have studied the differences between our subcorpora with growing n-grams, moving from bigrams up to heptagrams.

[4] We also report t-test and p-value for each pair of distributions. We bolded the statistics that reject H0 at the 0.05 significance level.

	ST	PT	t-test	p-value
FICTION	11.41	12.74	**-5.6**	<.001
ESSAY	10.54	13.73	**-14.2**	<.001
POPSCI	10.20	10.50	**-1.6**	<.001
INSTR	8.59	9.63	**-5.2**	<.001
SHARE	12.65	13.23	-0.5	0.5
SPEECH	10.08	9.83	1.2	0.2
TOU	10.22	12.34	**-9.04**	.001
ALL	11.12	14.39	**-2.45**	0.01

Table 2: Perplexity of the German-trained model on ST and PT.

Analysis of Pattern Diversity

Figure 1 illustrates the number of unique PoS n-grams used in the different registers of our German corpora by professionals (left graph) or students (right graph) – on the x-axis – as compared to the number of unique PoS n-grams used in the same registers by comparable German originals – on the y-axis.

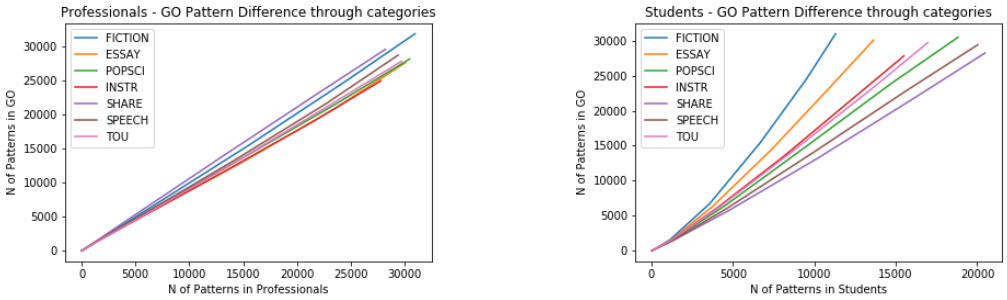


Figure 1: Differences between PoS n-grams going from bigrams to heptagrams.

Professionals tend to have register-specific variations that are substantially similar to those of the equivalent originals, while students appear to be less diverse than both comparable originals and professionals. This shows that the reason for the lower perplexity scores of the PoS-based language models for student translations is that students reuse fewer but more predictable structures. Professionals are more creative in their sentence structures: they are thus more perplexing for a general German model.

4 Conclusion

Our results show that, against our expectations, professional translations elicit higher perplexity scores from the target language model than students' translations. An analysis of the distribution of PoS patterns across registers shows that this apparent paradox is the effect of higher stylistic diversification and register sensitivity in professional translations. Our results contribute to the understanding of human translationese and shed light on the variation in texts generated by different translators, which is valuable for translation studies, multilingual language processing, and machine translation.

References

Biber, D., S. Conrad, and R. Reppen (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Hansen-Schirra, S., S. Neumann, and E. Steiner (2012). *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Berlin, New York: de Gruyter.

Kruger, H. and B. van Rooy (2012). Register and the Features of Translated Language. *Across Languages and Cultures* 13(1), 33–65.

Lapshinova-Koltunski, E. (2013). VARTRA: A Comparable Corpus for Analysis of Translation Variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria, 77–86. Association for Computational Linguistics.

Straka, M. and J. Straková (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Toury, G. (2004). Probabilistic explanations in translation studies: Welcome as they are, would they qualify as universals? In A. Mauranen and P. Kujamäki (Eds.), *Translation Universals: Do They Exist?*, Benjamins translation library, 15–32. J. Benjamins Publishing Company.

Cutting semantic corners? Patterns of lexical simplification in interpreting vs. translation

Yuri Bizzoni, Heike Przybyl, Elke Teich
Saarland University
yuri.bizzoni@uni-saarland.de, heike.przybyl@uni-saarland.de, e.teich@mx.uni-saarland.de

1 Introduction

Translation and Interpreting studies, especially when corpus-based, have accumulated a large body of empirical evidence that translations tend to share a number of features that distinguish them from non-translations, generally known as *translationese* and *interpretese*. Such features are usually divided into categories such as cross-linguistic interference (shining-through), normalization/standardization, ex- and implicitation, intensification, and simplification (Teich, 2003; Baroni and Bernardini, 2005; Zanettin, 2013; Xiao and Dai, 2014; Castagnoli, 2016; Pácelat, 2018).

Simplification is the general tendency of rendering the original text in a simplified manner in the target language. Simplification can happen in many ways: making shorter and more concise sentences; avoiding complex syntactic structures that are present in the source text; making explicit the boundaries of syntactic units by punctuation; or, from a lexical point of view, using a more general term instead of a more specific one (e.g. *entity* vs. *company*, *issue* vs. *question*), etc.

In simultaneous interpreting, simplification may be considered a producer-oriented strategy to cope with cognitive load (Gile, 2009; Kohn and Kalina,1996; Setton, 1999), one effect being choice of more general words.[5] Due to severe time pressure, it can be assumed that interpreters sometimes opt for a more general term whereas translators are able to think about or look up the exactly matching term. Therefore, it may be expected that interpreted speech exhibits more pronounced simplification effects regarding lexis than translation.

Previous studies on simplification in interpreting compared to translation using traditional corpus-based measures such as type token ratio (TTR), lexical density (LD), list heads or core vocabulary have shown mixed results. While translations are shown to be overall more simplified compared to their source texts (Laviosa, 1998), interpreted speech does not exhibit a clear pattern (Kajzer-Wietrzny, 2015). Instead, lexical density seems to be higher in interpreted than in original speech while at the same time interpreted speech is shorter on word level compared to the source speech (Russo et al., 2011; Kajzer-Wietrzny, 2015). However, there are exceptions to the observed trends and language combinations are found to have an influence (Dayter, 2018; Ferraresi et al., 2018). Such results are hard to interpret and possibly, there are methodological caveats (data may be too noisy or index hidden variables) or aggregated measures such as lexical density and TTR are simply not able to capture important differences between translation and interpreting at the lexical level.

In this paper, we take a different angle on simplification by turning to the semantic aspect of lexis and looking at patterns of semantic shift (Vinay et al.,1995). Focusing on nouns, we consider two semantic relations, abstract vs. concrete and specific vs. general, and apply selected computational methods and resources to assess semantic shifts from abstract to concrete and from specific to general.

[5] In production settings with less time pressure, simplification may also be an effect of audience design, i.e. linguistic adaptation to the communication conditions of some assumed recipients.

2 Approach and materials

We compare four datasets: English written originals and comparable German to English translations; English spoken originals, and comparable German to English interpreting transcripts.

The datasets belong to the same genre and register: they are transcriptions of European Parliament speeches by native speakers and their simultaneous interpreted renditions (EPIC-UdS, spoken) as well as the published written European Parliament speeches and their officially published translations (Europarl-UdS, written) (Karakanta et al., 2018).

Europarl-UdS			EPIC-UdS		
	sentences	words		sentences	words
ORG WR EN	372,547	8,693,135	ORG SP EN	3,623	68,548
TR EN DE	137,813	3,100,647	SI EN DE	4,080	58,218
ORG WR DE	427,779	7,869,289	ORG SP DE	3,408	57,049
TR DE EN	262,904	6,260,869	SI DE EN	3,622	59,100

Table 1: Corpus overview: Europarl-UdS (written) and EPIC-UdS (spoken). We compare the English subcorpora.

While level of concreteness and word generality are related concepts in linguistics, we resort to different resources to measure them independently. For concreteness, we use a human-annotated dictionary (Brysbaert et al., 2014) that reports the mean concreteness score of large amounts of English lemmas as assigned by several native speakers: values range from 5, indicating highly concrete terms (*yo-yo*, *tomato*) to 1, indicating highly abstract words (*conceptualistic*, *essentialness*).

For generality, we measured each noun's position in the hierarchical lexicon WordNet (Fellbaum, 2010) (e.g. *president* WN depth 12 - rated as specific vs. *people* WN depth 3 - rated as general). Since each noun can appear in different synsets, we compute its level of generality as the mean depth of the synsets in which it occurs.

Finally, we compare such measures with more classical measures such as standardized TTR and average word frequencies to gain a clearer profile of the lexico-semantic characteristics of our corpora.

3 Results

We make the following observations using the described methods and datasets (all measures were computed on a corpus-by-corpus basis):

1. *Standardized TTR*. Noun diversity is lower (higher NounTTR) both in translations and simultaneous interpreting when compared to originals in the same mode (p values ≤ 0.05).

2. There is no systematic difference between English original and translated texts in terms of nouns' average concreteness, nor in abstract/concrete ratios, if lemma frequency is not taken into account. If lemma frequency is instead taken into consideration, a systematic decrease in average concreteness between originals and translations can be detected. In other words, originals and translations draw from similar vocabularies, but translations seem to use fewer concrete nouns more frequently overall.

3. At the same time, the average concreteness of spoken data (English original and interpreted) is higher than the average concreteness of written data (English original and translated), making interpreting slightly more concrete than written translations, while translations and interpreting both are on average less

concrete than comparable originals. Both of these differences in concreteness, albeit relatively small, have proved statistically significant (p values ≤ 0.05). As a merely illustrative example, (1) shows the use of an abstract noun (a) in the written data as well as a concrete noun for spoken (b), with concreteness scores in brackets.

(1) a. translated: *we had some **difficulty** (1.9) settling in*
b. interpreted: *if you're talking about the sort of **housing** (4) that they require*

4. Written originals display the lowest high vs. low word frequency ratio, while interpreting transcripts display the highest. This can be seen as a complementary measure of TTR or lexical diversity, with written originals showing the largest ratio of rarely repeated words. Our written translation corpus shows more repetition than its comparable original, but less repetition than interpreting transcripts.

5. We use WordNet's depth to approximate word generality, considering terms with longer paths to the ontology's root as being more specific. With this approach, we cannot see significant differences between our datasets overall. Nonetheless, when only high frequency words are taken into consideration, they show a systematic skew towards generalization that goes from written originals (least general) to interpreting transcripts (most general). Example (2) below illustrates this phenomenon: while both sentences include nouns that are deep in the ontology, interpreting uses shallower nouns (e.g. *people*) with higher frequency. WordNet depth score in brackets.

(2) a. written original: *however, there were different **reactions** (5) to the **details** (5) of this new **vision** (8)*
b. interpreted: *there are **people** (3) who have more positive **visions** (8)*

4 Conclusion and outlook

Overall, our results point to symptoms of lexical simplification in interpreting for nouns in parliamentary speeches. We also show that a term's frequency plays an important role in detecting signs of simplification, since interpreting and translation appear to use more generic terms more often.

In our ongoing work, we seek to explain these findings by drawing on entropy and surprisal as indices of processing difficulty. Here, we may find that more general items come with a retrieval advantage (low entropy and surprisal), which may explain their higher frequency in interpreting (cf. Teich et al. (2020)). Also, we carry out analyses of other word classes, notably verbs and adjectives, and replicate the study for the German subset of the corpus in order to investigate effects of translation direction and language pair.

Acknowledgement

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

References

Baroni, M. and Bernardini, S. (2005). 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.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.4

Castagnoli, S. (2016). Investigating trainee translators' contrastive pragmalinguistic competence: a corpus-based analysis of interclausal linkage in learner translations. *The Interpreter and Translator Trainer*, 10(3):343–363.

Dayter, D. (2018). Describing lexical patterns in simultaneous interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM: International Journal of Interpretation and Translation.*

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Ferraresi, A., Bernardini, S., Petrović, M., and Lefer, M.-A. (2018). Simplified or not Simplified? the Different Guises of Mediated English at the European Parliament. *Meta*, 63(3):717–738.

Gile, D. (2009). *Basic Concepts and Models for Interpreter and Translator Training: Revised edition.* John Benjamins.

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

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

Kohn, K. and Kalina, S. (1996). The strategic dimension of interpreting. *Meta*, 41(1):118–138.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43(4):557–570.

Pácelat, I. L. (2018). Translation universals: Evidence from a study of Croatian and Italian translated texts. *CECL Papers*, page 104.

Russo, M., Bendazzoli, C., Sandrelli, A., and Spinolo, N. (2011). The European Parliament Interpreting Corpus (EPIC): Implementation and developments. *Linguistic Insights*, 147:53–90.

Setton, R. (1999). *Simultaneous Interpretation: A cognitive-pragmatic analysis*. John Benjamins.

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

Teich, E., Martînez Martînez, 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.

Vinay, J., Darbelnet, J., Sager, J., and Hamel, M. (1995). *Comparative Stylistics of French and English: A methodology for translation*. Benjamins Translation Library. John Benjamins Publishing Company.5

Xiao, R. and Dai, G. (2014). Lexical and grammatical properties of translational chinese: Translation universal hypotheses reevaluated from the Chinese perspective. *Corpus Linguistics and Linguistic Theory*, 10(1):11–55.

Zanettin, F. (2013). Corpus methods for descriptive translation studies. *Procedia -Social and Behavioral Sciences*, 95:20 – 32. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Cross-linguistic similarities in lexis: examining cognate activation through temporal and accuracy data from the Polish Interpreting Corpus (PINC)

Agnieszka Chmiel[1], Marta Kajzer-Wietrzny[1], Danijel Koržinek[2], Przemysław Janikowski[3]
Adam Mickiewicz University in Poznań[1], Polish-Japanese Academy of Information Technology, Warsaw[2], University of Silesia, Katowice[3]
achmiel@amu.edu.pl, kajzer@amu.edu.pl, danijel@pjwstk.edu.pl, przemyslaw.janikowski@us.edu.pl

The objective of this corpus-based study is to gain novel insight into the mechanism of lexical activation that underlie bilingual language control in conference interpreting on the basis of naturalistic data. We wanted to examine if and to what extent professional interpreters benefit from cross-linguistic similarities in lexis (whether they interpret cognates by applying their cognate or non-cognate equivalents).

Activation is one of the central mechanisms of bilingual control, largely examined on populations of bilinguals and manifested frequently through the cognate facilitation effect (Dijkstra, Van Hell, & Brenders, 2014). Such findings are easily explained by the Bilingual Interactive Activation Plus (BIA+) model (Dijkstra & van Heuven, 2002), recently extended into the Multilink model (Dijkstra et al., 2018), which posits language non-selective lexical access and a multiplicity of cross-level activations. As a result, seeing the English word *reduction* will activate the Polish cognate *redukcja* due to orthographic and phonological similarity. These words in turn will activate their shared semantic representation and the word *redukcja* will have a greater activation than *zmniejszenie*, the Polish word with a similar meaning but a different form.

The mechanism of lexical activation is largely in line with the theory of translation that suggests two strategies used in translation and interpreting (de Groot & Christoffels, 2006; Macizo & Bajo, 2006; Ruiz, Paredes, Macizo, & Bajo, 2008). One is vertical or conceptually mediated translation, according to which the source language message is first decoded and its conceptual representation is activated through phonological, morphological and semantic analysis and then its lexical representation in the target language is activated for production. The other is horizontal or structurally mediated translation in which source language utterances are directly transcoded into their target language equivalents thanks to memory associations (de Groot & Christoffels, 2006). This account is also in line with the literal translation hypothesis (Halverson, 2015) and the recursive model of translation (Schaeffer & Carl, 2013). Interpreters might use the horizontal or vertical strategy depending on context. We predict that if they use cognate rather than non-cognate equivalents when interpreting, especially in combination with a short ear-voice span (EVS), they opt for the horizontal strategy and use the most activated target language equivalent in accordance with the Multilink model. Alternatively, if they use non-cognate equivalents for cognates they favour the vertical strategy and use more inhibition to suppress the activated cognate equivalent. Oster (2017) found that cognate translations were less monitored and more frequent in oral than written translations. Defrancq (2015) reported that cognates interpreted by means of their cognate equivalents in the target language triggered a shorter EVS than other words elsewhere in the corpus. Hansen-Schirra, Nitzke, and Oster (2017) determined that the use of cognate translation equivalents in the written translation corpus was modulated by numerous factors, such as context, language status and translation experience. Our study extends this research by examining whether interpretation of cognates is modulated by word frequency and the number of translation equivalents and how EVS changes depending on the strategy used.

In line with Defrancq (2015), we predicted that cognates would generate a shorter ear-voice span (EVS) than non-cognate words since the former would enjoy facilitation. We also predicted a shorter EVS for cognate than non-cognate translations in line with the horizontal translation account. We expected

modulation of the EVS by such factors as frequency and the number of translation equivalents: we predicted shorter EVS for more frequent items and longer EVS for items with more translation equivalents for non-cognate translations and no effect for cognate translations.

In an attempt to test our predictions on naturalistic data, we created PINC (Polish Interpreting Corpus) – a new bidirectional, parallel and comparable time-annotated corpus of interpretations performed by professional conference interpreters in the European Parliament. It includes 190,000 words and is divided into four subcorpora: Polish original speeches, their English interpretations, English original speeches, their Polish interpretations. We identified cognate nouns represented in the subcorpus of Polish original speeches and aligned them with data on frequency (Mandera, Keuleers, Wodniecka, & Brysbaert, 2015), number of senses (taken from Maziarz et al. 2014), source text speed, interpretation speed and location in the source text. We then examined their translations in the parallel subcorpus of their interpretations into English. We evaluated translation accuracy and calculated ear-voice spans, or time lags between these words in the source and target texts.

Preliminary data shows the expected cognate facilitation effect in translation accuracy data only and not in temporal data. In other words, cognates were interpreted more accurately than non-cognates but there was no difference in EVS. Also, cognates were predominantly interpreted by means of their cognate equivalents, but again there was no difference in EVS generated by cognate and non-cognate equivalents. EVS was not modulated by frequency, but it was affected by the number of translation equivalents as words with more translation equivalents generated longer EVS. Our temporal data results are at a variance with Defrancq (2015) and do not support the Multilink model, which might be due to the different interpreting direction in that study and ours. However, our accuracy data do support the horizontal translation account. The lack of support for cognate facilitation in the EVS data might have various explanations. First, sentence context and language proficiency have been found to decrease cognate facilitation (Bultena, Dijkstra, & van Hell, 2014; Dijkstra et al., 2014; Libben & Titone, 2009): our participants were highly proficient in their working languages and processed cognates in sentence and text contexts. Also, the psycholinguistic models of cognate processing are largely based on visual presentation of stimuli, while our data comes predominantly from auditory presentation. Finally, we obtained few datapoints for non-cognate translations of cognates, which may have been insufficient to find the effect.

The study employs a corpus-based paradigm and shows support for the horizontal strategy used to interpret cognates in product data (accuracy) but not in process data (EVS). It thus contributes to our knowledge on how language control issues are managed by experienced interpreters in extreme conditions of high cognitive load and temporal constraints.

References

Bultena, S., Dijkstra, T., & van Hell, J. G. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *The Quarterly Journal of Experimental Psychology*, 67(6), 1214-1241.

de Groot, A. M. B., & Christoffels, I. K. (2006). Language control in bilinguals: Monolingual tasks and simultaneous interpreting. *Bilingualism: Language and Cognition*, 9(2), 189-201. doi:10.1017/s1366728906002537

Defrancq, B. (2015). Corpus-based research into the presumed effects of short EVS. *Interpreting*, 17(1), 26-45. doi:10.1075/intp.17.1.02def

Dijkstra, T., Van Hell, J. G., & Brenders, P. (2014). Sentence context effects in bilingual word recognition: Cognate status, sentence language, and semantic constraint. *Bilingualism: Language and Cognition*, 1-17. doi:10.1017/s1366728914000388

Dijkstra, T., & van Heuven, W. J. B. (2002). Modeling bilingual word recognition: Past, present and future. *Bilingualism: Language and Cognition*, 5(03). doi:10.1017/s1366728902283017

Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & RekkÉ, S. (2018). Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 1-23. doi:10.1017/s1366728918000287

Halverson, S. (2015). Cognitive translation studies and the merging of empirical paradigms: The case of 'literal translation'. *Translation Spaces*, 4(2), 310-340.
Hansen-Schirra, S., Nitzke, J., & Oster, K. (2017). Predicting cognate translation. In S. Hansen-Schirra, O. Czulo, & S. Hofmann (Eds.), *Empirical modelling of translation and interpreting* (pp. 3-22). Berlin: Language Science Press.
Libben, M. R., & Titone, D. A. (2009). Bilingual lexical access in context: evidence from eye movements during reading. *J Exp Psychol Learn Mem Cogn*, 35(2), 381-390. doi:10.1037/a0014875
Macizo, P., & Bajo, M. T. (2006). Reading for repetition and reading for translation: do they involve the same processes? *Cognition*, 99(1), 1-34. doi:10.1016/j.cognition.2004.09.012
Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behav Res Methods*, 47(2), 471-483.
Maziarz, M., Piasecki, M., Rudnicka, E., & Szpakowicz, S. (2014). PlWordNet as the cornerstone of a toolkit of lexico-semantic resources. In *Proceedings of the seventh global wordnet conference*: 304-312.
Oster, K. (2017). The influence of self-monitoring on the translation of cognates. In S. Hansen-Schirra, O. Czulo, & S. Hofmann (Eds.), *Empirical modelling of translation and interpreting* (pp. 23-39). Berlin: Language Science Press.
Ruiz, C., Paredes, N., Macizo, P., & Bajo, M. T. (2008). Activation of lexical and syntactic target language properties in translation. *Acta Psychol (Amst)*, 128(3), 490-500. doi:10.1016/j.actpsy.2007.08.004
Schaeffer, M., & Carl, M. (2013). Shared representations and the translation process: A recursive model. *Translation and Interpreting Studies*, 8(2), 169-190. doi:10.1075/tis.8.2.03sch

Contrastive evaluation of L1 and L2 translations based on the PIE method (Preselected Items Evaluation): a case study

Amy Colman, Winibert Segers, Heidi Verplaetse
KU Leuven
amy.colman@kuleuven.be, winibert.segers@kuleuven.be, heidi.verplaetse@kuleuven.be

Background

Translation into the foreign language – L2 translation (Stewart, 1999) – is widely discouraged by translation scholars. Newmark (1988/2001, p. 3) labels it "service translation", while Beeby Lonsdale (1996, p. 5) calls it "prose translation" and "inverse translation". International translation bodies also disapprove of the practice. The American Translators Association (ATA), for example, explicitly states that "professional translators work into their native language" (Durban, 2011, p. 16).

Nevertheless, L2 translation is common practice among professional translators. A survey conducted in 2014 by the International Association of Professional Translators and Interpreters (IAPTI) shows that more than half of all respondents (over 700 translators in all) translate into a language that is not their first language (IAPTI, 2015).

In addition, L2 translation is included in the curricula of most translator training institutions (Pokorn, 2016). While L2 translation in translator training has not been widely researched, most existing studies suggest that generally, the L2 translation output of students is of slightly inferior quality (Castillo Rodríguez, 2006; Pavlovic, 2007; Pokorn et al., 2019). However, it must be stressed that the differences between the translation output into the first language (L1) and the output into the second language (L2) cannot be easily quantified and compared objectively. After all, a major confounding factor is the use of different source texts, possibly of differing difficulty levels.

This highlights the need for further research into L2 translation in translator training, as well as an objective comparison between L1 and L2 student translations.

Research question and methodology

An experiment was designed to analyse the number and types of errors third-year students in applied linguistics make in L1 and L2 translation. Thirty students translated two texts: one from Dutch (L1) into English (L2) and one from English into Dutch. The target texts were evaluated by two evaluators. The first evaluator was the lecturer of the course during which the experiment was conducted and the second was a professional translator who is pursuing a PhD in Translation Studies. Due to the Covid-19 restrictions, the data was collected remotely. The students were provided with a written translation brief specifying the target audience, delivery format etc., and they were asked not to use machine translation.

To ensure ecological validity (Neunzig, 2011; Hansen, 2013), the students were presented with two source texts of approximately 250 words each on topics linked with the curriculum covered, namely shipping, trade and investment, to be translated in 90 minutes. The texts were selected jointly by the two evaluators.

Twenty items were preselected in each source text based on the curriculum and the intended learning outcomes. The students were unaware of which items had been chosen. The evaluation was based on

these items only, which were subsequently categorised using the error categories of the ATA Framework for Standardized Error Marking[6].

Before the experiment, the students were provided with a definition of L1 and they were asked to indicate, in writing in a brief questionnaire above the source text, whether this definition applied to them for Dutch. If not, they were asked to clarify how, where and when they acquired the language. Following each translation experiment, the participants indicated the perceived degree of difficulty of each text and the main difficulties they encountered. After the experiment, the students were asked which translation direction they found the most challenging.

The output of this experiment was evaluated using PIE (Preselected Items Evaluation), an item-based criterion- and norm-referenced analytical translation evaluation method (Kockaert & Segers, 2014; Segers & Kockaert, 2016; Kockaert & Segers, 2017; Segers et al., 2018; Tijtgat & Segers, 2019). PIE consists of five phases, viz., (1) the selection of a set number of items to be evaluated, (2) the dichotomous categorisation of translation solutions for each item as correct or incorrect, (3) the calculation of the test takers' raw scores, (4) the calculation of the item difficulty (p-value) and discrimination index (d-index) of each item, and (5) the calculation of the test takers' final scores based on the items with a good p-value and d-index (Segers & Kockaert, 2016). The first two phases are criterion-referenced, while the final two are norm-referenced.

PIE does not rely on the weighting of errors, but rather on the dichotomous categorisation of translation solutions. The evaluators jointly decide which translation solutions are correct and which ones are incorrect. This approach leaves no scope for a 'grey area', viz., a lack of consensus between evaluators during the evaluation. PIE also guarantees intra-rater reliability by eliminating the contrast and halo effect (Tijtgat & Segers, 2019).

Another advantage of PIE is that it allows for an objective comparison between the test takers' scores in the two translation directions through the calculation of the z-scores. These are standardised scores that enable evaluators to compare "scores on different kinds of variables by standardising the distribution" (McLeod, 2019). Standardising the scores eliminates the interference of the confounding factors, viz., the different source texts and translation directions. This provides an objective insight into the differences in the individual and collective student scores and rankings for L1 and L2 translation.

Aims of the study

This case study aims to gain a better insight into students' performance in the two translation directions. This could mark the first step in the development of a theoretical model for the objective evaluation and comparison of L1 and L2 translation in translator training. Despite its limited scope, this study may also provide valuable insights for the creation of contrastive corpora, including batteries of translation tests, for training purposes. Corpora can be created based on the insights obtained from the norm-referenced evaluation of the translation tests and the evolution of the students' scores during the academic year.

Preliminary findings

Surprisingly, the students achieved slightly higher scores in L2 translation. As for the types of errors made, they appeared to have struggled particularly with L1 spelling and L2 syntax. It is important to note that no logging software was used, so it is impossible to verify with certainty whether or not some students relied on machine translation, despite being instructed not to. The results of this experiment indicate, however, that L2 translation is not necessarily of inferior quality to L1 translation, despite what is often assumed in translator training.

[6] https://www.atanet.org/certification/aboutexams_error.php

References

Beeby Lonsdale, A. (1996). *Teaching Translation from Spanish to English: Worlds Beyond Words*. Ottawa: University of Ottawa Press.

Castillo Rodríguez, C. (2016). Translating Tourist Texts Into Non-Mother Tongue: An Experiment with a Multilingual Corpus, *Opción* (32), 419-436.

Durban, C. (2011). *Getting It Right. A Guide to Buying Translations*. Retrieved from https://www.atanet.org/publications/Getting_it_right.pdf.

Hansen, G. (2013). The translation process as object of research. In Millán, C. & Bartrina, F. (eds.), *The Routledge Handbook of Translation Studies* (pp. 88-101). Abingdon: Routledge.

IAPTI (2015). *Translation into a non-native language*. Retrieved from https://www.iapti.org/files/surveys/2/IAPTI_non-native_report.pdf.

Kockaert, H., & Segers, W. (2014). Evaluation de la traduction: la méthode PIE (Preselected Items Evaluation), *Revue Turjuman* 23(2), 232-250.

Kockaert, H. & Segers, W. (2017). Evaluation of legal translations: PIE method (Preselected Items Evaluation). *Journal of Specialised Translation,* (27), 148-163.

McLeod, S. A. (2019). Z-score: definition, calculation and interpretation. Retrieved from https://www.simplypsychology.org/z-score.html.

Neunzig, W. (2011). Empirical Studies in Translation: Methodological and Epistemological Questions. *TTR: traduction, terminologie, redaction*, 24(2), 15-39.

Newmark, P. (2001). *A Textbook of Translation*. Shanghai: Shanghai Foreign Language Education Press (Original work published 1988).

Pavlovic, N. (2007). *Directionality in Collaborative Translation Processes* (doctoral dissertation, Universitat Rovira i Virgili, Tarragona, Spain; University of Zagreb, Croatia). Retrieved from http://darhiv.ffzg.unizg.hr/id/eprint/2352/1/THESIS.pdf

Pokorn, N. (2016). Is it so different? Competences of teachers and students in L2 translation classes. *Rivista Internazionale di Tecnica della Traduzione / International Journal of Translation* 2016(18), 31-48.

Pokorn, N., Blake, J., Reindl, D., & Pisanski Peterlin, A. (2019). The influence of directionality on the quality of translation output in educational settings. *The Interpreter and Translator Trainer*, 1-21.

Segers, W. & Kockaert, H. (2016). Can Subjectivity be Avoided in Translation Evaluation? In Thelen M., van Egdom G-W., Verbeeck D., Lewandowska-Tomaszczyk B. (Eds.), *Translation and Meaning. New Series* (pp. 69-78). Frankfurt am Main: Peter Lang.

Segers, W., Kockaert, H. & Wylin, B. (2018). Vertaalevaluatie en subjectiviteit. *Tijdschrift n/f* (13), 41-51.

Stewart, D. (1999). Translators into the foreign language. Charlatans or professionals? *Rivista Internazionale di Tecnica della Traduzione 4*, 41-67.

Tijtgat, E. & Segers, W. (2019), Wat is een goede vertaling? Vertaalevaluatie: methodes en technieken. In *In balans. Een inleiding tot vertaal- en tolkwetenschap* (pp. 307-328). Leuven/Den Haag: Acco.

How the *PJ Masks* become "PJ Heroes"
A contrastive study of gender portrayal in the Dutch and Swedish dubbing of a popular cartoon

Reglindis De Ridder, Annika Johannson
Stockholm University
reglindis.deridder@nederlandska.su.se, annika.johansson@nederlandska.su.se

This study is part of a larger multimodal analysis project focussing on audiovisual translation for children. The broader aim is to establish how internationally distributed popular cartoons, such as *PJ Masks*, are localised for different markets. For this part of the project, both a qualitative and a quantitative linguistic corpus analysis are conducted, the results of which will be presented here. Such a systematic contrastive analysis (e.g. Doval and Sánchez Nieto 2019) aims to minimise the risk of confirmation bias or cherry picking only a small number of scenes. To this end, a trilingual parallel corpus of the English *PJ Masks* source text and both the Dutch (*Pyjamahelden*) and the Swedish (*Pyjamashjältarna*) audiovisual translations was built based on the transcripts of 12 episodes. In audiovisual fiction, information is transferred through images, sound, and language (Chaume 2012). As a result, the target audience's cognitive load when processing such fiction is substantial (Hvelplund 2018). In children's audiovisual fiction, relevant information conveyed through sound and images, therefore, often is explicitly rendered again through the spoken text to ensure all relevant information can be processed. This is another reason why a text corpus analysis in itself can be insightful, although of course it does not provide the full picture. The focus in the corpus analysis is the representation of the main characters through the linguistic channel. In the light of ongoing criticism of cartoons, as regards the lack of diversity (Götz et al. 2018) and gender stereotyping (Drottner 2018), this analysis of both source text and translations is concerned with possible changes in the gender representation that may occur in the audiovisual translation process. It aims to establish if, and if so to what extent, the characterisation and more specifically their gender portrayal changes in the translated products.

In Sweden, for instance, audiovisual translators are known to compensate for this gender stereotyping or the lack of diversity in the dubbing process of several imported cartoons (De Ridder 2019). Media scholar Kirsten Drotner (2018:384) posited that boys and girls are still typecast in similar ways in cartoons today, in that "boy characters are more likely than girl characters to be inventive, outgoing and problem-solving, while girl characters are more likely than boy characters to be attentive to relations and in need of assistance". From this, three categories were drawn for the contrastive analysis of the original *PJ Masks* and both audiovisual translations: problem solving, need for help, attention to relations. This qualitative and quantitative analysis of *PJ Masks* episodes examines if Drotner's statement holds true in this cartoon. It quickly became apparent that the Dutch translation deviates the most from the source text, while the Swedish translation remains closer to it. For this reason, utterances by all three heroes conveying such attention to relations, need for help, and problem solving have been analysed in the English source text and in the Dutch translation first. By way of illustration, in the category "need for help", we searched the corpus for utterances relating to asking for instructions (e.g. "OK, Catboy - what's the plan?"), asking for an explanation (e.g. "But how can a big train... disappear?"), explicitly asking for help (e.g. "I need help with my tail."), or admitting they're not in control (e.g. "Can't... hold... on! Agh!"). Subsequently, the number of instances of such utterances by the heroes in each category are counted and compared to check if there are significant differences in both language versions. We have observed, for instance, that explicit calls for help were not always conveyed in the Dutch translations. For example, in the source text the team leader says "I need help with my tail", which is translated in Dutch with "Ik zit hier nog steeds vast" [I'm still stuck here] omitting the explicit call for help, which was not the case in the Swedish translation. The use of the pronouns we/us was also systematically examined across all language versions and

characters to establish whether the heroes emphasised team work or their personal contributions. Here we found that the Swedish version contained even more instances of vi/oss [we/us] than the original English version. While in the Dutch translation significantly less instances of wij/we/ons [we/us] could be found.

The preliminary results of the linguistic analysis of the source text suggest that, in this cartoon, the female hero, counter to Drotner's claim, may, in fact, even be the least caring and the least in need of help, in that her male team members produce more utterances relating to both categories. Furthermore, she plays an equally substantial role in solving the problems with which the heroes are confronted in each episode. This is no different in the Dutch translation, as the quantitative data tallied in this translation more or less match the source text data. This seems to suggest that there are no significant changes in the character portrayal as revealed by the linguistic analysis of their utterances in both language versions. While the linguistic analysis of the source text may suggest that the female hero is less of a team player compared to her male team members, the sense of team spirit in general is significantly lower in the Dutch utterances across all characters and episodes because of this reduction in the wij/we/ons [we/us] pronouns. Still, it remains to be seen if this results in substantial changes in the gender depiction of the heroes, in general. To that end, a close analysis of translation shifts in the Dutch target text will be contrasted against the Swedish translation of the same lines to check if this reveals subtle or less subtle changes in the linguistic depiction of the PJ Masks.

Needless to say, such translation shifts in audiovisual translation for children and their effect on the representation of children in this multimodal product is highly relevant amidst ongoing criticism of gender and diversity issues in children's television. What is more, in times of ever more competing content providers importing and localising international productions to increase their content such audiovisual translations are worthy of closer scrutiny.

References

Chaume, Frederic. 2012. *Audiovisual translation: Dubbing.* New York: Routledge.

De Ridder, Reglindis. 2019. 'Het is verruktelijk'. Hoe audiovisuele vertalers het heft in eigen handen kunnen nemen. ['It is delicis'. How audiovisual translators can take matters into their own hands] *Filter. Tijdschrift over Vertalen*. 2019(4), pp.21-28.

Doval, Irene and Sánchez Nieto, María Teresa. 2019. *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins.

Drotner, Kirsten. 2018. Children and media. In *Mediated communication.* Ed. Philip M. Napoli, 379- 394. Berlin: De Gruyter.

Götz, Maya, Ole Hoffmann, Caroline Mendel, Dafna Lemish, Sebastian Scherr, Yuval Gozansky, Kirsten Huang, et al. 2018. Whose story is being told? Results of an analysis of children's tv in 8 countries. *Televizion.* 2018(31), pp.61-65.

Hvelplund, Kristian Tangsgaard. 2018. Eye tracking and the process of dubbing translation. In *Fast-forwarding with audiovisual translation.* Eds. Jorge Díaz Cintas and Kristijan Nikolić, 110-125. Bristol: Multilingual Matters.

Stable explanations in empirical translation studies: a cognitive-linguistic perspective

Gert De Sutter
Ghent University
gert.desutter@ugent.be

This paper contributes to the ongoing development in empirical translation studies towards more accurate, profound and encompassing explanatory models of translation behavior by integrating stable theoretical concepts from probabilistic linguistics, psycholinguistics and cognitive linguistics (thereby following the line of reasoning in Claes 2017). Several other scholars have already elaborated on the relevance of linguistic theorizing for translation studies (a.o. Alves et al. 2010, Steiner 2012, Halverson 2013), and this paper aims to add to that theory-aware development.

The introduction of stable theoretical concepts allows for more specific and interdisciplinary-valid research hypotheses that enables more precise predictions about translational outcomes, which can be tested empirically. As we have claimed elsewhere (De Sutter & Lefer 2020), corpus-based translation studies has a long tradition of post-hoc explanation: by means of rather vague translational mechanisms, such as explicitation, simplification and normalization, (corpus-based) translation scholars have attempted since the mid-1990's to make sense out of empirical results mostly *after* the research had been carried out. As a consequence, explanations are often tentative and sometimes contradictory, and it has caused theoretical and explanatory models in empirical translation studies underdeveloped (cf. also Halverson 2017).

In the present paper, we adopt well-studied and stable explanatory concepts from probabilistic linguistics, psycholinguistics and cognitive linguistics, operationalize them in a multifactorial research design, formulate specific research hypotheses and test these hypotheses by means of data culled from the Dutch Parallel Corpus (Macken et al. 2010). Our central research topic is a classic case of syntactic explicitation, viz. *zero/that* complementation in non-translated and translated English (from Dutch; the data were previously used in De Sutter & Vermeire 2020). All private, public and suasive complement-taking verbs (Quirk et al. 1985) which allow for optional *that* were extracted and manually verified; this resulted in a dataset containing 4,818 relevant instances, and were subsequently coded for presence vs. absence of *that* (response variable).

(1) He insisted that/Ø I repeat that after him. [dpc-gru-002593-en]

Instead of selecting explanatory variables that are hypothesized to have an effect on the choice between *that* and *zero* in a relative vacuum, as has often be done, we only select variables which can be related to three explanatory concepts from the linguistic fields mentioned above. These concepts, which are called domain-general principles in cognitive linguistics, are general principles that govern all aspects of language use (and not just syntactic variation) and are formulated *in accordance with* what is known about the functioning of the brain from other disciplines (this is called the *cognitive commitment*):

- Markedness of coding: constructions (including words, morphemes, syntactic structures…) that match the conceptualization of the source utterance (i.e. the 'meaning') best have a higher probability of being activated.
- Statistical preemption: when the representations of words and constructions are activated frequently together, this compositional entity becomes stored as a single node in the cognitive

network. This is called *entrenchment*. Such detailed constructions are activated faster than more schematic ones (i.e. constructions that need to be 'assembled' on the spot).

- Structural priming: language users tend to reuse recently activated constructions, independent of the specific lexical content of this construction (structural priming is considered a residual activation effect).

These three principles were operationalized by means of the following explanatory variables (following Kruger 2018):

(1) TenseModality {Present, Past, Modal, Non-finite}
(2) Aspect {Simple, Progressive, Perfect}
(3) Subject {Zero, Pronoun, Noun, Expl. It}
(4) Polarity {positive, negative}
(5) LemmaConstrFreq100Klog {continuous variable}
(6) Source-language structure {identical *dat*, other *dat*, no *dat*}

Variable (1)-(4) are proxies for the markedness-of-coding principle (with the values *Present*, *Simple*, *Zero/Pronoun* being the unmarked choices), variable (5) is a proxy for the statistical-preemption principle and variable (6) is a proxy for the priming principle. By testing which of these variables affect the choice significantly, and to what extent, we are able to verify the operation of these principles in translated language and to compare its operation with that in non-translated language. We therefore use a relatively new statistical technique called random forest modelling. This technique is related to conditional inference tree modeling, in which a response variable (*zero/that*) is predicted on the basis of a set of explanatory variables by recursively splitting the data in subsets using the predictor variable that is able to reach maximal homogeneity within each subset and maximal heterogeneity between the different subsets. After the first split, the process is repeated using the other predictor variables until no further split is able to significantly increase the homogeneity in each of the subsets. With random forests, one does not grow just *one* tree, but a complete forest of (for instance 3,000) trees, based on a random selection of predictors and data points, and then amalgamate the results over the entire forest. Preliminary results show that statistical preemption seems to play a major role in the *that* alternation, both in translated and non-translated English. Other cognitive principles play a marginal role.

On a general level, this paper hopes to show that empirical translation studies should not build its own theories independent of the theories in the neighboring fields; rather, it should borrow and adapt existing theoretical-linguistic insights to describe and explain translational phenomena such as explicitation. By doing so, it will also inform different strands in linguistics and psycholinguistics about the tenability of concepts and explanations, thereby creating a multi-directional relationship between both fields.

References

Alves, F., & Gonçalves, J. L. (2010). Relevance and translation. *Handbook of translation studies* 1, 279-284.

Claes, J. (2017). Probabilistic Grammar: The view from Cognitive Sociolinguistics. *Glossa. A Journal of General Linguistics* 2(1).

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.

De Sutter, G., & Vermeire, E. (2020). Grammatical Optionality in Translations: A Multifactorial Corpus Analysis of That/Zero Alternation in English Using the MuPDAR Approach. In L. Vandevoorde, J. Daems, & B. Defrancq (eds.) *New Empirical Perspectives on Translation and Interpreting*. New York & London: Routledge, 13-37.

Halverson, S. L. (2013). Implications of cognitive linguistics for translation studies. In A. Rojo & I. Ibarretxe-Antuñano (eds.) *Cognitive Linguistics and translation: Advances in some theoretical models and applications*. De Gruyter, 33-74.

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 methodological and theoretical traditions*. De Gruyter, 9-45.

Kruger, H. (2018). That again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*, 1-33.
Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch parallel corpus: A balanced copyright-cleared parallel corpus. META 56, 374-390.
Steiner, E. (2012). Generating hypotheses and operationalizations: The example of explicitness/explicitation. Cross-linguistic corpora for the study of translations: Insights from the language pair English-German, 55-70.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Cognitive load in simultaneous interpreting. Two theories held up against the light of corpus data.

Bart Defrancq, Koen Plevoets
Ghent University
bart.defrancq@ugent.be , koen.plevoets@ugent.be

Cognitive load is an evasive scape-goat in interpreting studies. Especially in simultaneous interpreting, it is blamed for errors, omissions and infelicities (EOR, Gile 2011), for disfluencies (Plevoets & Defrancq 2016; 2018), for the occurrence of specific lexical patterns (Ferraresi et al. 2017), for the build-up of stress in interpreters (Korpal 2016) and so on. However, the concept is poorly theoreticised and operationalised in interpreting studies. Gile's (2009) Efforts Model is still the most used functional model to analyse cognitive load and its different components. It describes interpreting as a complex balancing act in the management of cognitive resources. Seeber's (2011) Cognitive Load Model added to that an interference vector creating additional load. Both models present (simultaneous) interpreting as the performance of a multiple task involving a perceptual, a memory and a productive task, to which Gile adds resource management and Seeber a cognitive task, respectively. In Seeber's model, the perceptual and cognitive tasks are both split into sub-tasks dealing with the input and with the output. In both model tasks seem to consume indiscriminately the same amount of resources.

Operationalising cognitive load is challenging. Gile (2009) and Plevoets & Defrancq (2016; 2018) analyse output features; Seeber & Kerzel (2012) use pupillometry; Liang (2019) applies lexical simplification measures. There is a strong need to further analyse the features that these parameters pick up and to cross-validate the methods. For one, the immediacy of the targeted response to cognitive load is very different: pupillometry records exhibited increases less than 200 ms after the stimulus. Gile (2008), in contrast, envisages error responses beyond the sentence boundaries. The extent to which the different tasks contribute to cognitive load also varies from one study to another. In Seeber & Kerzel (2012) only one particular input feature is studied, namely verb-final structures in the source text that cannot be echoed in the target text. Plevoets & Defrancq (2016; 2018), in contrast analyse the contribution of a range of both input and output features.

This study builds on Plevoets & Defrancq (2016; 2018) in that it seeks to refine the one-dimensional accounts of input and output features by means of a systematic investigation of interactions between input and output features and their effects on the frequency of filled pauses in the output of simultaneous interpreters. Filled pauses are a widely recognised window on cognitive load in the psycholinguistic literature (Bortfeld et al. 2001). Data from the EPICG (Bernardini et al. 2018) were used to analyse the two-way interactions between delivery rate, lexical density, frequency of numbers, syntactic complexity and formulaicity.

The analysis consisted in an extensive search among all possible two-way interactions of these five features in the source text only, the target text only and between source and target text. Various models were tested in order to perform this search systematically and among all possibilities only three interactions were found to significantly contribute to an increased frequency of filled pauses. Moreover, they all the same feature in source and target text: input and output lexical density, input and output grammatical complexity and input and output formulaicity. As expected, low input formulaicity combined with low output formulaicity is associated with higher frequencies of filled pauses (Fig 1.). In contrast, and surprisingly, the interactions between input and output lexical density and input and output grammatical complexity play out differently. In both cases, it is a high input score combined with a low output score

that increases the number of filled pauses (Fig. 2 and 3). When both input and output scores are high for these features, interpreters do not show an increase in the number of filled pauses.

We can interpret these findings in two ways. The first explanation has to do with cognitive saturation or overload: when input load is (too) high, interpreters produce more disfluencies after which they deliver a reduced output. The second explanation sees the findings as arising from an interpreting strategy: when input load is high, interpreters actively try to reduce the output load by compressing and simplifying the information, but this reorganisation increases cognitive load and leads to disfluencies. The latter hypothesis has important theoretical implications. If substantiated, it implies that theoretical models of simultaneous interpreting need to include more components of linguistic processing than only input and output processing. Transforming dense or complex input into simpler or compressed output induces more cognitive load than keeping the parameters of input and output close. Seeber's (2011) Cognitive Load Model has an edge over Gile's (2009) Efforts Model in that respect, as it includes a third language processing component that the Efforts Model does not have. It would also support Lv & Liang's (2019) interpretation of lexical simplification as associated with high cognitive load.

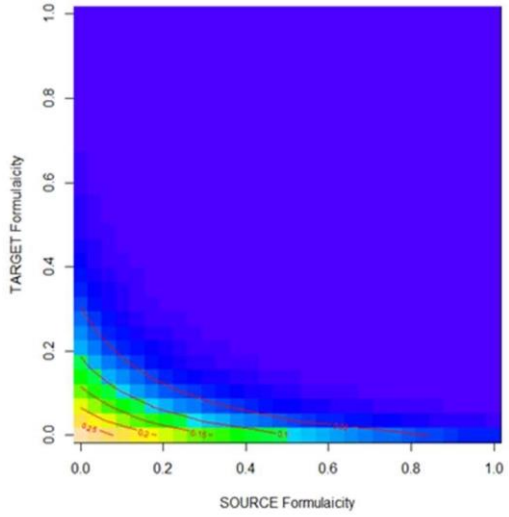


Fig. 1. Filled pause frequencies in interaction of input and output formulaicity

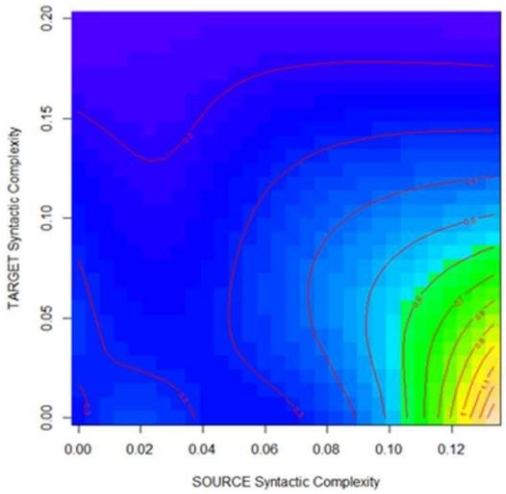


Fig. 2. Filled pause frequencies in interaction of input and output complexity

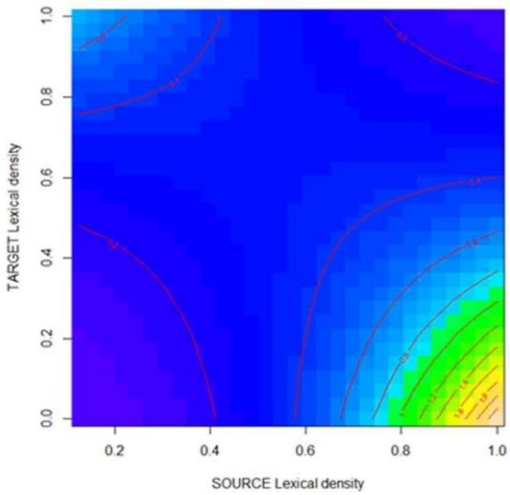


Fig. 1. Filled pause frequencies in interaction of input and output density

References

Bernardini, S., A. Ferraresi, M. Russo, C. Collard & B. Defrancq (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*. Singapore: Springer, 21-42

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F. & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech* 44, 123–147.

Ferraresi A. & M. Miličević (2017). Phraseological patterns in translation and interpreting. Similar or different? In G. De Sutter, M.-A. Lefer & I. Delaere (eds), *Empirical Translation Studies. New Methodological and Theoretical Traditions*. Berlin, Mouton-De Gruyter, 157-182.

Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum* 6, 59–77.

Gile, D. (2009). *Basic concepts and models for interpreter and translator training*. *Revised edition*. Amsterdam: John Benjamins.

Gile, D. (2011). Errors, omissions and infelicities in broadcast interpreting. Preliminary findings from a case study In A. Cecilia, A. Hild & E. Tiselius (eds.) *Methods and Strategies of Process Research: Integrative approaches in Translation Studies* pp. 201–218

Korpal, P. (2016). *Linguistic and psychological indicators of stress in simultaneous interpreting*. University of Poznan: PhD.

Qianxi Lv Q. & J. Liang (2019). Is consecutive interpreting easier than simultaneous interpreting? – a corpus-based study of lexical simplification in interpretation *Perspectives* 27(1), 91-107.

Plevoets, K. & B. Defrancq (2016). The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies* 11 (2), 202–224.

Plevoets, K. & B. Defrancq (2018). The cognitive load of interpreters in the European Parliament. A corpus-based study of predictors for the disfluency uh(m). *Interpreting* 20 (1), 1–29.

Seeber, K. (2011). Cognitive load in simultaneous interpreting: Existing theories – new models. *Interpreting* 13 (2), 176–204.

Seeber, K. & D. Kerzel (2012). Cognitive load in simultaneous interpreting: Model meets data. *International Journal of Bilingualism* 16 (2), 228–242.

BE verbs in a contrastive perspective: The case of *BÝT*, *BE* and *VÆRE*

Jarle Ebeling[1], Anna Čermáková[2], Signe Oksefjell Ebeling[1]
University of Oslo[1], University of Cambridge/Charles University[2]
jarle.ebeling@usit.uio.no, anna.cermakova@ff.cuni.cz, s.o.ebeling@ilos.uio.no

Drawing on data from the new International Comparable Corpus (https://korpus.cz/icc; Kirk et al. 2018), this paper reports on a cross-linguistic investigation of BE verbs in Czech, English and Norwegian, i.e. BÝT, BE and VÆRE. BE and its closest counterparts in other languages have been researched extensively (e.g. Bybee & Dahl 1989), however, to our knowledge, a detailed corpus-based contrastive mapping of the uses of the verbs in these three languages has not been performed before.

Etymologically the three verbs are related in a somewhat complex web of partly overlapping origins (cf. Rejzek 2015, OED Online 2019, Bjorvand & Lindeman 2019). It is therefore not surprising that, synchronically, there are both overlapping and non-overlapping uses, functions, and meanings. For example, all three verbs are used to express existence and they all function as auxiliaries to form the passive voice; see example (1). Moreover, they are used as linking verbs (copulas) with an adjectival complement, as in (2), but only English BE can be used as an auxiliary marking the progressive aspect (3a), while only Czech BÝT functions as an auxiliary to mark the past tense, as in (3b).[7]

(1a) The autumn evenings *were* marked by the Listowel races… (ICC-EN)
(1b) Je původní, potvrzuje dnešní majitel, který sem *byl* také před třiceti roky přenesen z Evropy. (ICC-CZ)
['It is original, confirms the contemporary owner, it *was* brought over here thirty years ago from Europe']
(1c) … brua *var* festet med store bolter. (ICC-NO)
['the bridge *was* secured with large bolts']

(2a) Facilities *were* rather spartan … (ICC-EN)
(2b) Vlak *je* zrezivělý, a tohle nástupiště *je* teď pusté. (ICC-CZ)
['The train *is* rusty and this platform *is* now deserted.']
(2c) Hun visste at hunden *var* adskillig reddere. (ICC-NO)
['She knew that the dog *was* considerably more scared']

(3a) Somewhere a baby *was* crying. (ICC-EN)
(3b) Rozsvítil *jsem* modrou lampičku a posadil se na posteli. (ICC-CZ)
['I turned on the blue lamp and sat up on my bed']

Against this backdrop, we wish to address the following research questions:

1. To what extent do these verbs overlap in meaning and use?
2. In the linking use, what kind of relationship does each of the verbs typically establish between the elements that are linked?

[7] Norwegian has an *-ende* form of the verb corresponding to the English *-ing* form, but it rarely combines with auxiliary VÆRE to form the progressive and we do not expect it to be attested in the material. However, with intransitive verbs, there are cases where VÆRE functions as an auxiliary to form the perfect aspect, as in: Det *er* blitt sent på natten. ['It *is* become late at night.'] According to the OED, this is also possible for BE in modern English, although it is "now largely replaced by *have* following the pattern of transitive verbs".

3. In relation to these verbs, what are the methodological challenges of cross-linguistic comparison of typologically different languages?

For the purpose of this study, we will extract our material from three of the already complete written components of the International Comparable Corpus: ICC-CZ, ICC-EN and ICC-NO. The investigation will be further restricted to one text type only: Creative writing.

In all three languages, these verbs are by far the most frequent ones: BÝT occurs 1,845 times, BE 1,735 and VÆRE 1,292 times in our data and, as expected, all have multiple meanings and functions. An analysis of a random sample of 100 concordance lines established two main, broadly comparable, uses across the languages: auxiliary and linking.

The samples show that the three languages differ somewhat regarding the proportion of the two major uses. Norwegian VÆRE is overwhelmingly linking in nature (95%), Czech BÝT has more auxiliary uses, with linking uses accounting for only around half of the occurrences in the sample. English BE is also predominantly linking (76%) in nature, but less so than Norwegian VÆRE. Table 1 shows the proportions of these categories in the three languages, with further sub-categorisations.

Table 1. The functions of BE, BÝT and VÆRE in ICC-/CZ/EN/NO

ICC-EN		**ICC-CZ**		**ICC-NO**	
Function	**#**	**Function**	**#**	**Function**	**#**
Linking	**76**	**Linking**	**52**	**Linking**	**95**
NP+BE+NP	**28**	NP+BÝT+NP	**15**	NP+VÆRE+NP	**31**
NP+BE+ADJP	**23**	NP+BÝT+ADJP	**17**	NP+VÆRE+ADJP	**24**
NP+BE+ADVP	6	NP+BÝT+ADVP	9	NP+VÆRE+ADVP	5
NP+BE+PP	9	NP+BÝT+ Ø	1	NP+VÆRE+PP	6
Linking, other	11	Linking, other	9	Linking, other	29
Aux	**24**	**Aux**	**47**	**Aux**	**5**
Prog Pass	12	Condit.	7	Pass	3
	12	Future	1	Perf	2
		Pass	6		
		Past	33		
Other	**0**	**Other** (idiom)	**1**	Other	**0**
TOTAL	**100**	**TOTAL**	**100**	**TOTAL**	**100**

Following this initial analysis, we will focus on the most frequent use across the three languages and carry out an in-depth analysis of the three verbs when they have a linking function on the basis of the complete dataset from ICC/Creative writing. An initial observation, based on the pilot study, is that, although the most frequent uses (NP+V+NP and NP+V+ADJP) seem to be similar across the languages, there are some interesting differences arising from the fact that the languages are typologically different, e.g. the many instances falling into the Linking, other category in Norwegian compared to English and Czech: 29 vs. 11 and 9 respectively.

In a more detailed analysis of the NP+V+ADJP pattern, we will semantically classify the adjectives in order to determine to what extent the three languages describe fictional subjects in the same way.

The methodological issues raised in research question (3) will be discussed in the light of the results emerging from the full-scale analysis of BE verbs in the three ICC components. Some of the potential challenges are already evident in the apparent mismatch of (sub-)categories shown in Table 1.

References

Bjorvand, H. & F.O. Lindeman. 2019. *Våre arveord. Etymologisk ordbok*. [3rd ed.] Oslo: Novus.

Bybee, J. L. & Ö. Dahl. 1989. The Creation of Tense and Aspect Systems in the Languages of the World. *Studies in Language* 13 (1), 51–103.

Kirk, J., A. Čermáková, S.O. Ebeling, J. Ebeling, M. Křen, K. Aijmer, V. Benko, R. Garabík, R. Górski, J. Jantunen, M. Kupietz, M. Šimková, T. Schmidt & O. Wicher. 2018. Poster presentation at the *UCCTS 2018 (UCCTS 2018 'Using Corpora in Contrastive and Translation Studies')* in Louvain-la-Neuve.

OED Online. 2019. "be, v.". Oxford University Press. https://www-oedcom.ezproxy.uio.no/view/Entry/16441?rskey=n7iZuu&result=4&isAdvanced=false (accessed February 06, 2020).

Rejzek, J. 2015. *Český etymologicý slovník*. [3rd ed.] Praha: Leda.

A Contrastive Rhetoric Analysis of Interactional Metadiscourse Markers in Online News: Articles Written in English and Arabic

Emna Fendri[1], Bochra Kouraichi[2]
Faculty of Arts and Humanities University of Sfax[1], Faculty of Humanities and Social Sciences University of Szeged[2]
Emna3000@yahoo.fr, kouraichi.bochra@stud.u-szeged.hu

Metadiscourse is recognized as a linguistic tool that is used by the writer to achieve his or her rhetorical purposes. It refers to the expressions "which explicitly organize a discourse or the writer's stance towards either its content or the reader" (Hyland, 2019, p. 16). It captures the relationship between the writer, the reader, and the text. Metadiscourse markers are used to project the participants in the text and to position them in the persuasive act. They are divided into two main categories: interactive and interactional markers. Interactive categories are said to "[h]elp to guide the reader through the text" while interactional markers "[i]nvolve the reader in the text". (Hyland, 2019, p. 58).

The focus of the present research is on interactional resources because they "focus more directly on the participants of the interaction" (Hyland & Tse, 2004, p. 170) by expressing the writer's stance and reader engagement in the text. In Hyland's (2019) model, interactional resources consist of hedges, boosters, attitude markers, self-mentions, and engagement markers. Hedges and boosters mark the writer's epistemic commitment to the content while attitude markers express his/her affective commitment to propositional content. Self-mentions reflect the writer's explicit presence in the text, whereas engagement markers are used to explicitly address the reader. Accordingly, the role of interactional metadiscourse markers (henceforth IMDMs) is of paramount importance in establishing an interpersonal relationship between the addresser and the addressee.

Online news articles are regarded as a source of informing and shaping public opinion. Interaction between the writer and the target reader is one way to strengthen the persuasiveness of the text. Persuasiveness is said to be dependent on culture and language according to the Contrastive Rhetoric approach (e.g. Kaplan, 1987; Connor, 2008). The present study aims to examine metadiscourse use as a persuasive strategy in English and Arabic. The type and frequency of IMDMs use will therefore be compared and contrasted between English and Arabic news articles to understand the way interpersonal relations are constructed across languages as a persuasive strategy.

Articles dealing with the same topic are collected from three different online news platforms that issue in both languages. A trending topic in spring 2021 is undoubtedly COVID-19 vaccinations and the tightening of protection measures across the globe. This is quite a sensitive topic that requires empathy and persuasiveness on the part of the journalists or news reporters; hence the significance of understanding the use of IMDMs in this context. The present study investigates the use of IMDMs in 15 English and 15 Arabic news articles to understand the nature of the writerreader relation across Arabic and English. The following research objectives are put forward: (i) to map out the use of IMDMs in both sub corpora, (ii) to compare and contrast the use of IMDMs in both sub corpora, (iii) to highlight what IMDMs use in the two languages reveal about the writerreader relationship, and (iv) to understand how IMDMs contribute to text persuasiveness in each language.

For this aim, a mixed-method approach was employed to analyze the data. A qualitative analysis is first carried out to identify IMDMs in both sub corpora. They are categorized into the five interactional subcategories (hedges, boosters, attitude markers, self-mentions, and engagement markers). The Text Inspector web tool was used to examine the English texts. The categories were then manually checked.

The Arabic articles were however analyzed manually since no automatic annotation was possible. Interrater reliability was then checked in the two sub corpora. Quantitative analysis was later carried out using the SPSS software. Repeated measures one-way ANOVA tests were computed to check whether the average use of IMDMs in the English and Arabic sub corpora were significantly different.

The analysis of the English corpus showed that hedges were most frequently used, followed by boosters then attitude markers, engagement markers and self-mentions. As to the Arabic corpus, boosters were more frequently used, followed by hedges, attitude markers, engagement markers and self-mentions. After a comparison of the quantitative results, the findings indicated that the use of IMDMs is statistically different between the two languages. Interpersonal relations are differently enacted across languages; more space is left for the reader to intervene in the persuasive act. It is manifested through a less frequent use of hedge in English articles. In Arabic articles, however, the more frequent use of boosters stifles the reader's voice and pushes him/her towards a specific interpretation. Besides, the analysis shows that the same markers may have different functions across the corpora. Qualitative analysis indeed foregrounds pragmatic differences between the two languages as to the use of IMDMs as a persuasive strategy. The similarity in the use of attitude markers, engagement markers and self-mentions in both subcorpora is explained in the light of the generic conventions of online news articles. The significance of the present study lies in the understanding of the rhetorical and pragmatic dimensions of persuasion in English and Arabic. Not only does the study help to unveil some mechanisms of intercultural communication through the use of metadiscourse, but it also contributes the to the understanding of this concept in Arabic as research is still limited in this area in comparison to English.

References

Bax, S. (2012). Text Inspector. Online text analysis tool. Available at: https://textinspector.com/.

Connor, U. (2008). Mapping multidimensional aspects of research: Reaching to intercultural rhetoric. In U. Connor, E. Nagelhout, & V.W. Rozyki, (Eds.), Contrastive rhetoric: Reaching to intercultural rhetoric (pp. 299-315). Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Hyland, K. (2019). Metadiscourse: Exploring Interaction in Writing. London, New York: Bloomsbury.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. Applied linguistics, 25(2), 156-177.

Kaplan, R. B. (1987). Cultural thought patterns revisited. In U. Connor, & R.B. Kaplan (Eds.) Writing across languages: Analysis of L2 text (pp. 9-21). The United States: AddisonWesley Publishing Company, Inc.

Collocations in English/Italian translation: 'chiaro/scuro, 'bright/dark'

Gioia Franchi, Daniel Henkel
Univ. Paris 8 Vincennes-St. Denis
gioia.franchi@gmail.com, daniel.henkel@univ-paris8.fr

1. Introduction

Collocations have been described as one of the "cornerstones" of corpus linguistics (Bernardini, 2007) but remain difficult to define and measure adequately (Evert, 2009), a difficulty which is compounded when they are studied in translation. This study focuses two pairs of approximate translational equivalents, It. 'chiaro/scuro' and En. 'bright/dark', with a view to determining whether collocations in translated texts are the same as in untranslated texts and to what extent they may be influenced by the source-language.

2. Methods

This study is based on an Italian-English corpus consisting of "comparable" and "parallel" (cf. McEnery & Xiao, 2007) subcorpora, in what Johansson (2007) describes as a "bidirectional" model. Public-domain literature was collected from Project Gutenberg and LiberLiber.it. The size of the corpus was limited by the smaller number of Italian works for which English translations were available (cf. Zanettin, 2002). In all, 40 works of the 19th and 20th centuries by 40 different authors/translators were compiled into four subcorpora comparable in terms of size and diversity:

Original English (Or.En) *10 works/10 authors* 1,236,778 tokens	**English-from-Italian (En.I)** *10 works/10 translators* 1,021,652 tokens
Italian-from-English (It.E) *10 works/10 translators* 1,258,290 tokens	**Original Italian (Or.It)** *10 works/10 authors* 1,062,137 tokens

Table 1. Characteristics of Italian and English subcorpora

Texts were tagged in TreeTagger, and a second fully lemmatized version was produced so that collocation strength could be calculated for word-types (lemmas) rather than word-forms. The 'Collocates' tool in AntConc 3.5.9 was used to extract collocations with a frequency of ≥3 co-occurrences and '+Log-likelihood ($p<0.05$)' activated in the 'Statistics' options to eliminate results based on insufficient data. Two queries were made for each lemma: one targeting left-hand bigrams ("l-bigrams"), and another targeting right-hand bigrams ("r-bigrams"). Collocations which obtained MI scores ≥3 in the bidirectional corpus were then filtered according to the number of texts. Bigrams found in only one text were considered as idiosyncratic and excluded. Finally, median frequencies were calculated for each subcorpus.

3. Results

3.1 Collocations

Only associations between lexical categories, i.e. noun, verb, adjective or adverb, are reported:

Original English	English-from-Italian	Italian-from-English	Original Italian
X+'bright' 'very'(3.95), 'so'(3.37); 'eye'(3.63)	X+'bright' 'so'(3.31)	X+'chiaro' 'abbastanza'(6.80), 'più'(4.75), 'bene'(3.82), 'molto'(3.66), 'così'(3.37); 'capello'(5.36), 'voce'(4.54), 'giorno'(4.20), occhio(3.3); 'apparire'(5.99), 'vedere'(3.04)	X+'chiaro' 'più'(4.39), 'così'(3.22); 'occhio'(4.89), 'voce'(4.11), 'giorno'(3.29); 'parlare'(4.07), 'vedere'(3.17)
'bright'+X 'ray'(8.51), 'vision'(7.63), 'colour'(6.68), 'flower'(5.64), 'blue'(5.60), 'eye'(5.21), 'hope'(5.15), 'light'(3.80), 'thing'(3.13)	'bright'+X 'blue'(7.15), 'eye'(4.62)	'chiaro'+X 'idea'(5.10)	'chiaro'+X 'occhio'(3.79)
X+'dark' 'very'(3.45), 'still'(3.02); 'thick'(5.72), 'tall'(5.36), 'great'(3.37); 'grow'(5.87)	X+'dark' 'quite'(4.85), 'rather'(4.73), 'almost'(4.20), 'very'(3.97); 'great'(3.19)	X+'scuro' 'più'(MI=3.34); 'castano'(MI=10.65), 'vestito'(MI=6.62), 'occhio'(MI=6.49), 'ombra'(MI=6.07)	X+'scuro' 'più'(3.48); 'occhio'(4.24)
'dark'+X 'complexion'(7.49), 'blue'(6.49), 'brown'(5.97), 'curl'(5.89), 'eye'(5.88), 'shape'(5.65), 'wood'(5.25), 'corner'(5.17), 'field'(5.10), 'hair'(4.86), 'shadow'(4.82), 'figure'(4.76), 'sky'(4.60), 'night'(4.45), 'street'(4.39), 'against'(3.76), 'water'(3.73), 'hour'(3.71), 'side'(3.31)	'dark'+X 'passage'(7.13), 'corner'(5.77), 'blue'(5.32), 'hair'(5.25), 'room'(4.88), 'street'(4.65), 'eye'(4.12), 'hour'(4.03), 'night'(3.93), 'face'(3.22)	'scuro'+X (none)	'scuro'+X (none)

Table 2 Collocate bi-grams and MI scores.

3.2 Frequencies

Original English		English-from-Italian		Italian-from-English		Original Italian	
bright	2.23	bright	0.88	chiaro	1.72	chiaro	2.21
dark	3.55	dark	3.65	scuro	1.07	scuro	0.69

Table 2 Median frequencies per 10,000 words (f/10k)

4. Discussion

Due to the small size of the corpora, these results do not provide a complete picture of the distribution of 'bright/dark', 'chiaro/scuro' in English and Italian. Nonetheless, certain tendencies are worth noticing.

Although they have practically identical frequencies in Or.En and Or.It, 'bright' and 'chiaro' are not used systematically as translations of one another, since both occur less frequently in translation, especially 'bright' (f=2.23/10k vs. f=0.88/10k). As for 'dark/scuro', in Or.En 'dark' is much more frequent (f=3.55/10k) than 'scuro' in Or.It (f=0.69/10k). Indeed, in Or.It the quasi-synonyms 'oscuro' and 'cupo' also have median frequencies of f=0.65/10k and f=0.68/10k respectively, and the preference for one or another varies with different authors. In Italian-translated-from-English, translators overall favor 'scuro' (f=1.07/10k vs. f=0.87/10k for 'oscuro' and f=0.38/10k for 'cupo'). Given the relatively small size of the corpus, the fact that the same semantic field is split between three quasi-synonyms, with fewer occurrences for each individually, helps explain why 'scuro' on its own has no collocations in Or.It, while translators' preference for 'scuro' helps account for the greater number of collocations found in It.E.

Qualitatively, the collocations found in translation are similar to those in Original English and Italian. The association with 'eye' and 'occhio' is common to all four lexemes in all four subcorpora. In Or.En., 'dark' is used to describe other human features: 'complexion', 'curl', 'hair', and spatial entities: 'wood', 'corner', 'field', 'street'. Many of these collocations can be found in English-translated-from-Italian as well. In Or.It, 'chiaro' forms collocations with 'occhio', 'voce' and 'giorno' all of which can be found in Italian-translated-from-English, even though 'giorno' and 'voce' do not correspond to any collocates of 'bright' in English. The only two collocations for 'bright' in English-translated-from-Italian, 'bright+blue' and 'bright+eye', are present in Original English, the other collocations for 'bright' in Or.En, however, are lost in translation. On the whole, translators appear to follow the same stylistic conventions as authors, but the influence of the source language is perceptible as well.

The strongest interlinguistic influence apparent in these results is negative: 'chiaro' and 'scuro' have few or no collocations in Original Italian, and presumably this is reflected in the much smaller number of collocations in English-translated-from-Italian compared with Or.En (2 vs. 9 for 'bright'+X, 10 vs. 19 for 'dark'+X). A certain degree of positive “shining-through” (Teich, 2003) can be seen as well, insofar as 'chiaro' and 'scuro' have a few more colocations in Italian-translated-from-English than in Or.It. Most noticeably, 'castano+scuro' and 'ombra+scuro' reflect the influence of the English collocations 'dark+brown' and 'dark+shadow'.

The collocations of 'bright' and 'chiaro' reveal certain differences in their respective semantic fields: with 'voce', 'chiaro' is used to describe an acoustic perception, while 'bright' takes on a metaphorical interpretation with 'hope'. Such differences help explain why these two supposed translational equivalents are not in fact used as translations of one another as often as might be expected. The reasons why 'chiaro', with a similar frequency to 'bright', has fewer collocations in Or.It. remain to be elucidated, while for 'scuro' a larger corpus will be needed to obtain a more complete distributional profile.

References

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3): 209-226.

Bernardini, S. (2007). Collocations in translated language: Combining parallel, comparable and reference corpora. In Fourth Corps Linguistics conference held at the University of Birmingham (pp. 27-30).

Evert, S. (2009). Corpora and collocations. In Corpus linguistics. An international handbook, vol. 2. Berlin: Mouton de Gruyter.

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

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

McEnery, T., & 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.

Zanettin, F. (2002). CEXI: designing an english Italian translational corpus. In Teaching and Learning by Doing Corpus Analysis (pp. 327-343). Brill Rodopi.

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

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

CBTS meets readability research: New methodological insights for the study of the simplification hypothesis

Thomas François, Marie-Aude Lefer
UCLouvain
thomas.francois@uclouvain.be, marie-aude.lefer@uclouvain.be

Ever since the publication of Laviosa's (1998a, 1998b) pioneering work, the study of lexico-syntactic simplification has held centre stage in corpus translation research concerned with the typical features of translated text (see e.g. Corpas Pastor et al. 2008, Grabowski 2013, Kajzer-Wietrzny 2015, Ferraresi et al. 2018, Lv & Liang 2019). Lexico-syntactic simplification can be defined as translators' tendency to produce target texts that are less informationally dense, less lexically varied and/or sophisticated, and less syntactically elaborate than comparable texts in the same language that have been produced in *unmediated* circumstances, i.e. in situations of monolingual text production (cf. Bernardini et al. 2016: 64-65). While empirical evidence of simplification has been found in different translated languages, including non-European ones, and translation modalities (written translation, consecutive and simultaneous interpreting), methodological advancement in the area has been rather modest. To date, corpus translation scholars have mainly relied on Laviosa's linguistic operationalizations of simplification (lexical density, core vocabulary coverage, list head coverage and mean sentence length) without providing aggregate simplification profiles of translated texts. In the present paper, we propose an innovative approach to the study of simplification in translation that aims to move beyond this set of operationalizations. Specifically, the approach draws on insights from readability research, which has recently undergone major advances under the influence of machine learning and natural language processing (NLP) (cf. Benjamin 2012, Collins-Thompson 2014). This paradigm offers robust, sophisticated analytical models with which to investigate the simplicity/complexity spectrum in language. In particular, NLP-informed readability studies rely on a wide range of simplification parameters, which are more likely to capture text dimensions that might be overlooked by shallow parameters, and make use of advanced statistical methods to aggregate these parameters.

In this study, we use the translated and original proceedings of the European Parliament as a test case. More specifically, we rely on two French subcorpora of Europarl-*direct* (Cartoni and Meyer 2012): (i) an original French subcorpus made up of 1,880 speeches delivered by 192 speakers (630,000+ tokens) and (ii) a French-translated-from-English subcorpus containing 5,257 speeches given by 237 speakers (1.5 million+ tokens). Both subcorpora have been POS-tagged with the TreeTagger (Schmid 1995). One of the reasons for choosing Europarl is its availability and multilingualism, which make it possible to enlarge the empirical foundation of the approach presented here by replicating it on other Europarl datasets. The simplification analyses are based on François's (2011) readability model, which includes both classic and NLP-enabled simplification parameters. For the present analyses, we have selected the 19 most relevant lexical, syntactic and discursive parameters from François's set (see Table 1). The selection has been guided by previous research in both CBTS and NLP-informed readability studies. We have chosen parameters that have proved useful in previous readability research and that may be meaningful in translation research, alongside more traditional simplification indicators *à la* Laviosa.

We apply a twofold approach in the statistical analysis. First, drawing on Laviosa's methodology, we analyze all variables separately in order to detect simplification effects at the variable level. First, we use Wilcoxon rank sum tests to compare the means of each parameter in the original French (OF) and translated French (TF) conditions. In addition, to better characterize the size of the effect of translation on simplification, we compute point-biserial correlation coefficients between the two conditions (OF and TF) and each simplification parameter. Second, we use multivariate analyses so as to consider the effect of

translation on all the simplification parameters at once, offering a more comprehensive way of testing the simplification hypothesis. In particular, we apply a principal component analysis (PCA) transformation of the 19 parameters, which allows us to deal with multicollinearity issues. With the help of the PCA, we reduce the analysis to the three most explanatory components, which together explain 61% of the variance. The first component encodes lexical frequency information, the second is a mix of lexical and syntactic information and the third corresponds to word length. In view of the fact that the three components are independent, we apply a Wilcoxon rank sum test on each component to compare the two conditions (OF and TF).

Table 1: Lexical, syntactic and discursive simplification parameters examined in the study

Parameter type	Simplification parameter	Description
Lexical	TTR-L	Lexical variety index 1: ratio of the number of types to the number of tokens (based on lemmas)
	NormTTR-L	Lexical variety index 2: type-token ratio, based on lemmas and normalized per 100 words
	LEX/GRAM	Lexical density index 1: ratio of lexical words to grammatical words
	LEX/ALL	Lexical density index 2: ratio of lexical words to all running words
	ConcDens	Estimate of conceptual density as defined by Kintsch et al. (1975) and computed with *Densidées* (Lee et al. 2010)
	MWL	Mean word length: average number of letters per word
	PW10	Proportion of words of 10 letters or more
	Syllper100	Number of syllables per 100 words (see François and Miltsakaki 2012)
	CVC200 CVC1000 CVC2000 CVC5000	Core vocabulary coverage: percentage of lemmas found in the top-frequency list extracted from the web-crawled frTenTen reference corpus (cf. Jakubíček et al. 2013). We calculated the variables on the basis of four different list sizes: top 200, 1000, 2000 and 5000 most frequent lemmas in frTenTen
	GMLF	Geometric mean of lemma frequencies
	75LF	75th percentile of the probability distribution of lemmas per speech
	90LF	90th percentile of the probability distribution of lemmas per speech
Syntactic	MSL	Mean sentence length: average number of words per sentence
	%LongSent	Percentage of sentences that are longer than 30 words (cf. Daoust et al. 1996)
Discursive	PRO/NAM	Ratio of pronouns to proper names
	PRO/NOM+NAM	Ratio of pronouns to nouns (common nouns and proper names)

The results show that the simplification hypothesis is largely confirmed: translated texts are found to be simpler, both lexically and syntactically, than original texts. Lexically, for instance, translations are less dense (LEX/GRAM, LEX/ALL), contain fewer words of 10+ letters (PW10) and rely more on high-frequency words (all CVC variables, with CVC5000 being the most powerful parameter). As regards syntax, we find that sentences are shorter in translations (MSL) and that there are fewer sentences of 30+ words in translations (%LongSent). Interestingly, the two pronoun-noun ratios we used point to discursive complexification in translation. This trend might be linked to a well-known cross-linguistic

contrast between French and English, French being more nominal than English. In the presentation, we will zoom in on a few key methodological takeaways for the study of simplification in CBTS.

References

Benjamin, R.G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* 24(1): 63-88.

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

Cartoni, B. & Meyer, Th. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, May 2012, Istanbul, Turkey.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165(2): 97-135.

Corpas Pastor, G., Mitkov, R., Naveed, A., et al. (2008). Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*. (AMTA2008: The Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, 21-25 October 2008). Stroudsburg: Association for Machine Translation in the Americas, 75-81.

Daoust, F., Laroche, L., & Ouellet, L. (1996). SATO-CALIBRAGE: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue Québécoise de Linguistique* 25(1): 205-234.

Ferraresi, A., Bernardini, S., Milicevic Petrovic, M., & Lefer, M.-A. (2018). Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta* 63(3): 717-737.

François, Th. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Doctoral dissertation, unpublished. Louvain-la-Neuve: Université catholique de Louvain.

François, Th. & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In: *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations* (PITR 2012) (Montreal, Canada, June 7, 2012), 49-57.

Grabowski, Ł. (2013). Interfacing corpus linguistics and computational stylistics. Translation universals in translational literary Polish. *International Journal of Corpus Linguistics* 18(2): 254-280.

Jakubicek, M., Kilgarriff, A., Kovar, V., Rychly, P., & Suchomel, V. (2013). The TenTen corpus family. In: *7th International Corpus Linguistics Conference*, 125-127.

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

Kintsch, W., Kozminsky, E., Streby, W., McKoon, G., & Keenan, J. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior* 14(2): 196-214.

Laviosa, S. (1998a). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4): 557-570.

Laviosa, S. (1998b). The English Comparable Corpus: a Resource and a Methodology. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (eds.) *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome Publishing.

Lee, H., Gambette, Ph., Maille, E., & Thuillier, C. (2010). *Densidées* : calcul automatique de la densité des idées dans un corpus oral. In *RECITAL'2010 : 12ième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*. (Montréal, Canada, 19-22 July 2010).

Lv, Q. & Liang, J. (2019). Is consecutive interpreting easier than simultaneous interpreting? – a corpus-based study of lexical simplification in interpretation. *Perspectives* 27(1):91-106.

Schmid, H.L. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. (Dublin, Ireland).

A contrastive study of English *finally/eventually*, Dutch *eindelijk/uiteindelijk* and French *finalement/enfin*

Lobke Ghesquière, Gudrun Vanderbauwhede
University of Mons
lobke.ghesquiere@umons.ac.be, gudrun.vanderbauwhede@umons.ac.be

In Germanic and Romance languages, adverbs that express the end of a sequence or the conclusion of a long (thought) process are quite a challenge for translators. Deciding whether to translate Dutch *eventually* as French *finalement* or *enfin* is not an easy task. Similarly, choosing between the English forms *eventually* and *finally* can be tricky. Dictionaries are not very helpful as the definitions of these words are highly similar or even identical and the adverbs are often listed as (near-)synonyms. Also, there is the added difficulty, especially for student translators, that FR *éventuellement* and DU *eventueel* are false friends for EN *eventually* which lacks the possibility meaning of the former. Despite their remarkable crosslinguistic morphosyntactic and pragmatic-semantic similarity, to our knowledge no thorough comparative study of these adverbs has been carried out so far. This paper aims to go some way in filling this gap and reports on a synchronic study of six of these adverbial markers – EN *finally/eventually*, DU *eindelijk/uiteindelijk* and FR *finalement/enfin*.

Whereas the electronic *Algemene Nederlandse Spraakkunst* ('General Dutch Grammar') and the *Van Dale Woordenboek Nederlands* ('Van Dale Dutch Dictionary') describe both DU *eindelijk* and *uiteindelijk* as conclusive discourse markers and consider them synonyms, *Taaladvies.net* (language advice website of the Dutch Language Union) states that *eindelijk* no longer functions as a conclusive discourse marker, having become obsolete in the sense of 'in time' or 'at the latest, finally'. Whereas *uiteindelijk* introduces the last element in an enumeration, *eindelijk* is now only used in the meaning 'after a long wait, after a lot of effort', adding a notion of impatience or relief to the sentence.

A somewhat similar division of labour is suggested for EN *eventually* and *finally*. Whereas both *finally* and *eventually* can be used to signal the final element in a temporal sequence ('in the end', 'after a long time or a lot of effort'), only *finally* has a more textual function, introducing the last (and crucial) point or conclusion of a discourse (OED s.v. *eventually*; OED s.v. *finally*; Quirk *et al.* 1985: 1454, 1471). For *finally*, the semantic difference is said to be reflected in a positional preference, with temporal uses favouring mid-sentence position and textual ones sentence-initial position (Swan 2016: 465).

The *Trésor de la langue française informatisée* (TLFi) offers a detailed description of FR *finalement* and *enfin*. Both adverbs are said to signal the conclusion of an enumeration or a thought process, with *enfin* being more polysemic as it also has several affective meanings. The French adverbs are thus considered partly synonymous, while, in our experience, native speakers often seem to have a rather clear preference for either *finalement* or *enfin* and tend to correct non-native speakers when they make the 'wrong' choice.

Contrastive grammars (e.g. van Baardewijk-Rességuier – van Willigen-Sinemus 2001, Henn *et al.* 2004) mention the adverbs' polysemic nature and their possible translations, but there is no extensive literature on the subject.

The *Van Dale* bilingual dictionaries simplify things as follows for Dutch and French: DU *eindelijk* expresses the temporal meaning 'after a long wait' and is translated by FR *enfin*, while DU *uiteindelijk* functions as a conclusive discourse marker and is translated by FR *finalement*. Whereas the preliminary results of our DU-to-FR corpus study confirm that DU *eindelijk* is mostly translated by *enfin* in French, the correspondence *uiteindelijk – finalement* does not seem to hold and needs to be looked into. Translations

from French into Dutch show a preference for the translation of FR *finalement* by DU *uiteindelijk*, but the translation of FR *enfin* seems to be less straightforward, with *eindelijk, uiteindelijk, ten slotte* and *tot slot* as its most common translations in Dutch.

The *Van Dale* bilingual EN/DU dictionary mentions *uiteindelijk* but not *eindelijk* as a possible translation for both *finally* and *eventually*, but favours *tenslotte* 'in closing'. This seems to be confirmed by the data, at least for the textual, sentence-initial uses of *finally* which are almost exclusively translated as *tenslotte*. The *Van Dale* DU/EN dictionary suggests *finally* as the translation for both *eindelijk* and *uiteindelijk*, but mentions *eventually* as an alternative. The translation data actually show a very diverse picture including many different alternative translations, omissions and paraphrases that will need to be looked into to understand the different strategies used by the translators.

For EN/FR the *Collins* and *Cambridge* dictionaries suggest only *finalement* as the corresponding form for *eventually*. For *finally*, they suggest both *finalement* and *enfin* for the temporal use. For the conclusive use of *finally* they suggest only *enfin* and the alternative forms *pour finir/terminer*. This is very similar to what was suggested for EN/DU. For FR/EN, both dictionaries only suggest *finally* and not *eventually* as translations for both *finalement* and *enfin*.

In this study, we will draw up typologies of the different uses of these adverbial markers and compare them qualitatively and quantitatively, focussing on their semantics/pragmatics and their structural behaviour. Monolingual corpora will be used to get a clear idea of the full range of uses of the individual items studied. Translation data will allow us to assess the degree of intertranslatability of the constructions as well as to come to a better understanding of the different language-internal uses. The translation data for this study are exhaustive extractions from the 10-million-word bi-directional *Dutch Parallel Corpus* (DPC), comprising Dutch, English and French texts with Dutch as a pivotal language. The monolingual DU, FR and EN data will be drawn from *DBNL*, *Frantext* and *WordbanksOnline* (usbooks and brbooks subcorpora) respectively. Random samples of 120 tokens for each adverb were extracted using a simple word query. All hits will be analysed in terms of syntactic position, meaning, collocational preferences, and, if applicable, translation and translation strategy.

It is our aim to *finally* sharpen the now sometimes blurry distinction between these adverbial forms, both language-internally and cross-linguistically, and *eventually* perhaps even come to concrete tips for student translators.

Corpora

DBNL: Digitale Bibliotheek voor de Nederlandse Letteren [Digital Library of Dutch Literature]. Available online at http://www.dbnl.org.
DPC: Dutch Parallel Corpus. Available online at https://www.kuleuven-kulak.be/dpc/conc.
Frantext: Available online at http://www.frantext.fr.
WB: Wordbanks Online. Available online at https://wordbanks.harpercollins.co.uk.

References

Cambridge Dictionary. Available online at https://dictionary.cambridge.org.
Collins Dictionary. Available online at https://www.collinsdictionary.com.
E-ANS: *elektronische Algemene Nederlandse Spraakkunst* [General Dutch Grammar]. Available online at http://ans.ruhosting.nl/e-ans/.
Henn, C., Vromans, J. & Bijleveld, H.-A. (2004[2]). *Pratique du néerlandais de A à Z*. Brussels: Didier Hatier.
Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English Language*. London: Longman.
Swan, M. (2016[4]). *Practical English Usage*. Oxford: OUP.
Taaladvies.net. Available online at https://taaladvies.net.
TFLi: *Trésor de la Langue Française informatisé*. Available online at http://atilf.atilf.fr.

van Baardewijk-Rességuier, J. & van Willigen-Sinemus, M. (2001[4]). *Matériaux pour la traduction du néerlandais en français*. Bussum: Coutinho.
Van Dale Online. Available online at https://www.vandale.nl.

Learner L2 translation corpus as a resource for translator trainers in facilitating the development of trainees' phraseological competence

Justyna Giczela-Pastwa
University of Gdańsk
justyna.giczela@ug.edu.pl

For almost 30 years now, corpora have been effectively and efficiently used by translation scholars, practising translators and translator trainees: comparable corpora have been used to compare various aspects of language use interlinguistically and interculturally, parallel ones – to identify particular translation shifts and strategies. More recently, yet another type of corpus, i.e. learner corpus, has attracted increasing attention of translation researchers. As it is expected, analysing considerable collections of translations done by trainees gives a well-informed insight into the actual problems that are faced and struggled with, and allows trainers to better incorporate appropriate themes into training. Presumably, it is Uzar and Waliński (2001/2007), as well as Bowker and Bennison (2003), who reported on the first attempts to systematically examine translator trainees' output gathered in the form of corpora, although the latter called them Student Translation Archives. The next two decades witnessed a growth in the number of projects designed to analyse learner translation corpora (Castagnoli et al. 2011; Espunya 2013; Wurm 2013, 2016; Kutuzov and Kunilovskaya 2014; Castagnoli 2016; Martínez and Vela 2016; Fictumova et al. 2017; Granger and Lefer 2017, 2018, 2020; Oțăt and Vîlceanu 2018; Alfuraih 2019), intended to answer a whole array of research questions, concerning — among other things — fluency, accuracy, quality, types of errors, translation strategies, and contrastive pragmalinguistic competence.

It is envisaged that the proposed study may contribute to the aforementioned line of research, by offering insight into the output produced by translator trainees who are non-natives of the target language (TL). The research material consists of L2 translations (Polish into English) prepared by 20 students on the MA in Specialised Translation Programme, University of Gdańsk, of three subsequent cohorts (2018-2021). On average, each student's contribution to the corpus amounts to over 2500 tokens and 10 translations, which makes the corpus a collection of over 50 000 tokens and approximately 200 texts. They predominantly pertain to topics of contract law, commercial law, and property law, with a few dealing with criminal law issues. The corpus is balanced in terms of translation conditions (Bowker and Bennison, 2003: 105), i.e. it includes both translations done for a graded assessment, and translations done in the final exam in the programme.

As already signalled, L2 translation lies at the core of the project. Although it has been strongly questioned and until recently marginalised in translation-related research, in the case of languages of limited diffusion, such as Polish, L2 translation is widely practised and often unavoidable. Similarly, L2 translation training is usually only briefly mentioned in training materials. This shortage, together with research-informed comments on the imperfect quality of L2 translator trainees' output (e.g. Heaney 2016; Pontrandolfo 2016; Huertas Barros and Castro 2018; Orlando 2018), validates investigation into L2 translation performance by trainees. The proposed type of exploration may be of particular value to trainers who themselves are non-native speakers of a TL. It has been suggested that a training environment in which learners and teachers move in the same direction along the native / non-native axis may have certain advantage over that in which the positions and outlooks of learners and teachers are dissimilar (Stewart 2008; Pokorn 2009; Hagemann 2019). As far as languages of limited diffusion are concerned, similarly as is the case with the translation marketplace, it is often unavoidable to welcome training under a supervision of a non-native speaker of the TL. Therefore, in some contexts trainers cannot rely on the invaluable native speaker's intuition. It seems then reasonable to resort to other resources in order to better assist learners in their development of L2 translation skills. Certainly, among such resources are learner translation

corpora, which precisely indicate what lowers the levels of accuracy and adequacy achieved by trainees in their translations.

Thus, the aim of the research is threefold. On the one hand, the primary intention is (a) to identify areas of difficulty that could later be dealt with more thoroughly in class; on the other hand, the investigation is (b) to recognize good practices that, while adopted by the trainees, enhance the overall level of textual fit in their translations. Last but not least, the objective of the project is (c) to explore which types of interference from the trainees' L1 are the most perceptible.

The methodology draws from the multilingually comparable corpus method (Hansen-Schirra and Teich 2008), more explicitly referred to as the comparable-parallel corpus method (Biel 2016), previously used in a project focused on L2 Polish-English legal translation (Giczela-Pastwa 2019). The present analysis has been carried out with the use of Sketch Engine (Kilgariff et al. 2014). The compiled learner L2 translation corpus (LL2TraC) is analysed against a self-collected reference corpus, consisting of (1) L2 English translations of Polish legal acts, done by professionals and brought out by three different publishers (over 1.6m tokens), and three ready-made corpora available through Sketch Engine, i.e. Eur-Lex English 2/2016, British Law Report Corpus and enTenTen20. In addition to the monolingual comparable corpora, a parallel corpus of the Polish source texts is used for consultation purposes, in order to study L1 interference. The alignment of the source texts and the translations was carried out with the use of LF Aligner 4.21. Each of the twenty subcorpora of LL2TraC (consisting of all the translations performed by the same trainee), as well as an additional subcorpus that consists of single translations by all of the trainees, are searched for keywords. Next, a set of the most salient keywords is singled out for further scrutiny. It consists in examining (multi-)word sketches, sketch differences and bilingual word sketches that contain the keywords and corresponding source language lexical units, in order to identify marked patterns that lower the textual fit of the translations. Additionally, by consulting the parallel corpus, an attempt is made to discover a probable stimulus for the observed markedness.

The analysis focuses particularly on markedness at the level of non-terminological bundles / bundles with low termness. It has been observed that translator trainees tend to use unidiomatic combinations, unsmooth expressions and inappropriately marked phrases in the TL (Heaney 2016: 83; Pontrandolfo 2016; Huertas Barros and Castro 2018: 47; Orlando 2018: 33). The scrutiny of collected data is targeted at identifying marked patterns in trainees' translations (in terms of frequency or structure), and determining the reasons for the occurrence of such untypical collocations.

References

Alfuraih, R. F. (2019). The Undergraduate Learner Translation Corpus: A new resource for translation studies and computational linguistics. *Language Resources & Evaluation*. https://doi.org/10.1007/s10579-019-09472-6

Biel, Ł. (2016). Mixed Corpus Design for Researching the Eurolect: A genre-based comparable-parallel corpus in the PL EUROLECT project. In E. Gruszczyńska & A. Leńko-Szymańska (eds.) *Polskojęzyczne korpusy równoległe / Polish-language Parallel Corpora*. Warszawa: Instytut Lingwistyki Stosowanej, 197–208.

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

Castagnoli, S. (2016). Investigating Trainee Translators' Contrastive Pragmalinguistic Competence: A corpus-based analysis of interclausal linkage in learner translations. *The Interpreter and Translator Trainer* 10(3), 343–363.

Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N. & Volanschi, A. (2011). Designing a Learner Translation Corpus for Training Purposes. In N. Kübler (ed.) *Corpora, Language, Teaching and Resources: From theory to practice*. Bern: Peter Lang, 221–247.

Espunya, A. (2013). The UPF Learner Translation Corpus As a Resource for Translation Training. *Language Resources & Evaluation*. https://doi.org/10.1007/s10579-013-9260-1

Fictumova, J., Obrusnik, A. & Stepankova, K. (2017). Teaching Specialized Translation: Error-tagged translation learner corpora. *Sendebar* 28, 209–241.

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*. Abingdon: Routledge, 48–65.

Granger, S. & Lefer, M.-A. (2017). Bridging the Gap between Learner Corpus Research and Translation Studies: The multilingual student translation corpus. 4th Learner Corpus Conference, Bolzano.

Granger, S. & Lefer, M.-A. (2018). MUST: A collaborative corpus collection initiative for translation teaching and research. In S. Granger, M.-A. Lefer & L. Aguiar de Souza Penha Marion (eds.) *Book of Abstracts: Using corpora in contrastive and translation studies conference* (5th edition). Louvain-la-Neuve: Centre for English Corpus Linguistics/Universite Catholique de Louvain, 72–73.

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

Hagemann, S. (2019). Directionality in Translation and Revision Teaching: A case study of an a–b teacher working with b–a students. *The Interpreter and Translator Trainer* 13(1), 86-101.

Hansen-Schirra, S. & Teich, E. (2008). Corpora in Human Translation. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics. An International Handbook*, Vol. 2. Berlin: Walter de Gruyter, 1159–1175.

Heaney, D. (2016). A Comparison of Conventional Metaphors on the Euro Crisis: A systematic approach for specialised L2 translation training. In G. Garzone, D. Heaney & G. Riboni (eds.) *Language for Specific Purposes: Research and translation across cultures and media*. Newcastle upon Tyne: Cambridge Scholars Publishing, 80–100.

Huertas-Barros, E. & Buendía Castro, M. (2018). Analysing Phraseological Units in Legal Translation: Evaluation of translation errors for the English-Spanish language pair. In S. Goźdź-Roszkowski & G. Pontrandolfo (eds.) *Phraseology in Legal and Institutional Settings: A Corpus-based interdisciplinary perspective*. Abingdon: Routledge, 41–60.

Kilgariff, 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.

Kutuzov, A. & Kunilovskaya, M. (2014). Russian Learner Translator Corpus: Design, research potential and application. In P. Sojka, A. Horák, I. Kopeček & K. Pala (eds.) *Text, Speech and Dialogue: 17th International Conference*. Brno, Proceedings. Cham: Springer, 315–323.

Martínez, J. M. & Vela, M. (2016). SubCo: A learner translation corpus of human and machine subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association, 2246–2254.

Orlando, D. (2018). The Problem of Legal Phraseology: A case of translators vs lawyers. In S. Goźdź-Roszkowski & G. Pontrandolfo (eds.) *Phraseology in Legal and Institutional Settings: A Corpus-based interdisciplinary perspective.* Abingdon: Routledge, 27–40.

Oțăt, D. & Vîlceanu, T. 2018. Learner Corpora and Cross-Linguistic Applications. *Scientific Bulletin of the Politehnica University of Timişoara. Transactions on Modern Languages* 17 (1), 5–18.

Pokorn, N. 2009. Natives or Non-Natives? That Is the Question… teachers of translation into language B. *The Interpreter and Translator Trainer* 3(2), 189–208.

Pontrandolfo, G. 2016. La evaluación en el aula de traducción jurídica. Una experiencia de análisis de errores en la combinación español-italiano. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, 29(1), 296–331.

Stewart, D. 2008. Vocational Translation Training into a Foreign Language. *inTRAlinea* 10. http://www.intralinea.org/archive/article/1646

Uzar, R. & Waliński, J. 2007 (2001). Analysing the Fluency of Translators. In W. Teubert (ed.) *Text, Corpora and Multilingual Lexicography*. Amsterdam: John Benjamins, 135–145.

Wurm, A. 2013. Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE). *trans-kom* 6(2), 381–419.

Wurm, A. 2016. Presentation of the KOPTE Corpus. Version 2. https://www.academia.edu/24012369/Presentation_of_the_KOPTE_Corpus_and_Research_Project

Corpus-based translation studies: Can we do better? Insights from a combined quantitative and qualitative survey

Sylviane Granger, Marie-Aude Lefer
UCLouvain
sylviane.granger@uclouvain.be, marie-aude.lefer@uclouvain.be

Over a quarter of a century after the emergence of corpus-based translation and interpreting studies (Baker 1993, Shlesinger 1998), it seems worth while to look back and take stock of the most recent developments in the field. In this presentation, we offer a thorough review, both quantitative and qualitative, of recent corpus studies of translation and interpreting with a view to describing their key characteristics in terms of corpus data, corpus-linguistic methods and research foci, identifying potential gaps in research to date and suggesting promising avenues for the future. More precisely, we present a research survey based on 186 corpus studies published in English in twelve top-rated translation and interpreting journals (*Across Languages and Cultures*, *Babel*, *Interpreting*, *inTRAlinea*, *Journal of Specialised Translation*, *Meta*, *Perspectives*, *Target*, *The Interpreter and Translator Trainer*, *trans-kom*, *Translation & Interpreting* and *Translation and Interpreting Studies*) between 2012 and 2019, thereby covering a period of eight years. The corpus studies included in the survey all meet two basic requirements: they rely on machine-readable corpora, i.e. electronic collections of texts, and they make use of corpus-linguistic techniques and tools to analyse them (e.g. concordancing, keyword extraction). Other recent surveys of translation and interpreting studies, such as Candel-Mora & Vargas (2013), Zanettin et al. (2015) and van Doorslaer & Gambier (2015), mostly rely on bibliometric records (titles, abstracts and keywords). These surveys provide some useful information on corpus use in translation and interpreting research and have the advantage of covering publications in a wide range of languages. However, the method on which they are based has its limitations, the main one being that bibliometric records fail to provide detailed information on many key features of corpus-based translation and interpreting studies (e.g. corpus methodology). In contrast to these studies, the present survey relies on an in-depth manual exploration of the full texts of the 186 research articles included in our dataset.

The survey provides a wealth of insights into the current status of the field, testifying to the growing maturity of corpus research in translation and interpreting studies while also identifying areas where progress has been relatively modest. First, it reveals that the genuine corpus studies represent 11% of the total number of articles published in English in the 12 selected journals. This finding ties in with Zanettin et al. (2015: 12), which shows that corpus-based studies accounted for c. 7% of translation and interpreting research in 2011. The fact that this percentage is higher than that established by the authors for 1997 (c. 3%), coupled with our own average proportion of 11% for the 2012-2019 period, suggests that corpus-based translation and interpreting studies are experiencing an upward trend which reflects the growth of corpus linguistics in general (Liao & Lei 2017: 4). It is important to point out, however, that there are significant differences between the journals (e.g. 29% of corpus-based studies in *Across Languages and Cultures* vs 5% in *The Interpreter and Translator Trainer*). We have classified the articles included in our survey dataset into three main categories, according to their main research foci and objectives: empirical, methodological-theoretical and applied. The empirical category includes corpus studies that are primarily focused on description and devoted to specific linguistic phenomena (e.g. grammar, lexis and terminology, discourse and pragmatics, semantics) and translation features (e.g. explicitation, normalization, simplification, convergence). The methodological-theoretical category subsumes three main types of contribution: calls for methodological and theoretical advancement, such as proposals for the adoption of methods and theories borrowed from neighbouring disciplines, literature reviews and overviews, and descriptions of new corpora and corpus tools for translation and interpreting studies. The applied category covers four major types of corpus application in translation and interpreting

studies, namely corpus use in translator and interpreter training, professional practice (language industry), translation quality assessment, and machine translation. In the presentation, we focus on empirical studies, which account for approximately two thirds of our dataset, and describe the main emerging trends. In particular, we analyse in detail a range of specific aspects, such as research focus (linguistic focus and translation features), corpus design (corpus types, corpus size, modality and register, languages investigated), corpus techniques (basic vs advanced) and use of statistics. As regards corpus design, for instance, the survey reveals that parallel corpora are used twice as frequently as monolingual comparable corpora, a finding that runs counter to Baker's (1995) call to move away from source text-target text comparisons.

One of the survey's most important findings concerns the use of corpus techniques and statistics. The analysis shows that the majority of the empirical studies rely on fairly basic techniques (frequency, concordancing), which were already promoted in Baker's early papers. More advanced techniques are found to be far less frequently used. Our survey also shows that most studies rely on simple descriptive statistics (such as relative frequencies) or monovariate inferential statistics, although advanced corpus techniques and elaborate statistical testing have recently started to gain momentum, following several calls for methodological rigour in the field (cf. De Sutter et al. 2012). The presentation ends with some forward-looking suggestions for the field.

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*. Amsterdam/Philadelphia: Benjamins, 233-250.

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target* 7(2), 223-243.

Candel-Mora, M.A. & Vargas-Sierra, C. (2013). An Analysis of Research Production in Corpus Linguistics Applied to Translation. *Procedia* 95, 317-324.

De Sutter, G., Goethals, P., Leuschner, T., & Vandepitte, S. (2012). Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures* 13(2), 137-143.

Liao, S. & Lei, L. (2017). What We Talk about When We Talk about Corpus: A Bibliometric Analysis of Corpus-related Research in Linguistics (2000-2015). *Glottometrics* 38, 1-20.

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

van Doorslaer, L. & Gambier, Y. (2015). Measuring Relationships in Translation Studies. On Affiliations and Keyword Frequencies in the Translation Studies Bibliography. *Perspectives* 23(2), 305-319.

Zanettin, F., Saldanha, G., & Harding, S.-A. (2015). Sketching Landscapes in Translation Studies: A Bibliographic Study. *Perspectives* 23(2), 161-182.

Interpreters' explicitating styles: A corpus study

Ewa Gumul, Magdalena Bartłomiejczyk
University of Silesia
ewa.gumul@us.edu.pl, magdalena.bartlomiejczyk@us.edu.pl

Idiosyncratic features of language use have attracted the attention of scholars for a long time. In Translation Studies, the question of style is particularly complex, as the authorship of translated texts is shared between the original author and the translator. Traditionally, the emphasis was on the former and the translator was seen as, ideally, an "invisible" transmitter (Venuti 1995). In literary translation, his/her success was therefore mainly measured by how faithfully s/he managed to reflect the author's style (cf. Baker 2000: 244). Baker (2000) and Saldanha (2011) take a radically different approach, proposing methodologies to examine what remains constant as one and the same translator works with a number of source texts representing various authorial styles. Both also test their proposals on small corpora containing literary translations by two individuals in each case. Comparing literary translators pairwise has actually become a standard line of inquiry, and this approach has spread to Interpreting Studies, where researchers contrast "lean" and "abundant" styles (see, e.g., Van Besien and Meuleman 2008, Baxter 2019).

There is some controversy as to which features might be treated as the best indicators of style, and, in practice, most researchers focus on the ones accessible to corpus linguistics tools, e.g. the type-to-token ratio or optional *that* after reporting verbs (Gumul 2017: 254, see also Rybicki 2012). Interpreting may be particularly suited for analysing style, because, unlike in the case of written translation, we undoubtedly deal with utterances of a single individual, free of unidentifiable input from proofreaders, editors, etc. Nevertheless, in contrast to the translator's style, the interpreter's style has rarely been researched.

We propose to explore interpreters' styles on the basis of our recently compiled EP-Poland corpus. It is a bidirectional parallel corpus of EP plenary texts, containing all Polish and English contributions to eleven debates held between January 2016 and February 2020, devoted mainly to the rule of law crisis in Poland triggered by its populist Law and Justice government. The total size of the corpus is over 157,000 tokens and about 20 h 45 min of recordings counting both source texts and target texts. English-Polish and Polish-English interpreting is represented almost evenly.

To be able to focus on individual interpreters, we first had to identify them in the recordings by the timbre of their voice. This has been done semi-automatically, using the technology based on the X-vector method (Synder et al. 2018) and defining new voice samples 'by ear'. Within the whole corpus, we identified 36 interpreters, out of whom, however, some interpret only 1-2 speeches. We therefore needed to establish a threshold for inclusion into the style analysis. Considering that plenary speeches are predominantly very short (2 min 16 sec on average), we settled on 15 minutes (counting source text material only) and at least four different source language speakers. 15 interpreters meet these criteria, and their outputs jointly account for 7 h 34 min 45 sec, that is, about 73% of the whole subcorpus of target texts.

As a first step, we investigate these interpreters' explicitating styles, in line with Gumul (2017). Murtisari defines explicitation as "shifts of meaning from the implicit to the implicit or simply to higher degree of explicitness" (2013: 332). In line with this definition, explicitation is understood in our study as a shift from source-text implicitness to an explicit rendering in the target text or a shift from an explicit encoding to a more explicit one through focus, emphasis, or lexical choice. The scope of such shifts ranges from cohesion-related surface additions or specifications (adding organising items, intensifying cohesive ties, lexicalising pro-forms, filling out elliptical constructions), through syntactic transformations (replacing

nominalisations with verb phrases) to other texture-enriching shifts (adding modifiers and qualifiers, inserting hedges, including explanatory remarks, disambiguating lexical metaphors). Gumul argues that "[e]xplicitation may be considered as an ideal indicator of a translator's or interpreters style because it is a feature which is optional and which by its nature is absent from the source text. This allows to filter the source-text variable and the influence of systemic differences […]" (2017: 257). As most forms of explicitation (e.g., disambiguation of lexical metaphors, addition of explanatory remarks, etc.) are not accessible to corpus linguistic tools, instances of explicitation are annotated manually with a view to establishing whether explicitation is interpreter-specific and to what extent individual interpreters may diverge from one another in this respect.

In our analysis we employ some of the explicitating styles identified in Gumul's (2017) study, which were distinguished taking into account the parameters of frequency and consistency. The first parameter of frequency has been operationalised by means of three main styles: lean, medium, and abundant – reflecting, respectively, scarcity, moderation, and frequent use of explicitation. In terms of consistency, two additional styles were identified: strategic and circumstantial. The first one describes interpreting outputs characterised by a consistent pattern of recurrent shifts, while the second refers to those renditions in which interpreters depart from their default settings due to the current interpreting circumstances and the difficulties they experience while processing the text under constraints of simultaneous interpreting.

The results reveal a wide variety of explicitating styles among the analysed 15 interpreters. This finding implies that explicitation in simultaneous interpreting of professionals is a highly idiosyncratic behaviour. Moreover, given the specificity of the SI task, explicitation in this mode of interpreting is not only conditioned by the so-called default settings of a given interpreter and his or her overall interpreting style, but is also to a large extent shaped by the unpredictability of the explicitating pattern in constrained conditions. Speaker-specific problems, such as fast delivery rate or deficits in coherence, tend to change the habitual explicitating style of an interpreter.

References

Baker, M. (2000). Towards a methodology for investigating style of a literary translator. *Target* 12(2), 241–266.

Baxter, R. N. (2019). Style versus strategy in simultaneous interpreting: Different approaches and their effects. *Quaderns. Revista de Traducció* 26, 287–305.

Gumul, E. (2017). *Explicitation in simultaneous interpreting : A study into explicitating behaviour of trainee interpreters.* Wydawnictwo Uniwersytetu Śląskiego.

Murtisari, E. T. (2013). A Relevance-based framework for explicitation and implicitation in translation. An alternative typology. *Trans-kom. Journal of Translation and Technical Communication Research* 6(2), 315–344.

Rybicki, J. (2012). The great mystery of (almost) invisible translator: Stylometry in translation. In M. Oakes & M. Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, 231–248

Saldanha, G. (2011). Translator style: Methodological considerations. *The Translator* 17(1), 25–50.

Snyder, D., D. Garcia-Romero, G. Sell, D. Povey & S. Khudanpur. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5329–5333.

van Besien, F. & C. Meuleman (2008). Style differences among simultaneous interpreters. *The Translator* 14(1), 135–155.

Venuti, L. (1995). *The translator's invisibility*. London: Routledge.

Seem and *appear* and their Norwegian verbal counterparts: a cross-register contrastive study

Hilde Hasselgård
University of Oslo, ILOS
hilde.hasselgard@ilos.uio.no

The meanings of verbs such as seem and *appear* combine features of evidentiality with epistemic modality (e.g. Aijmer 2009, Lampert 2011, Mortelmans 2014). The verbs occur in a variety of constructions, such as copular, as in (1), catenative, as in (2) and clause-introducing, with or without a dummy subject, as in (3). The examples come from the English-Norwegian Parallel Corpus (ENPC), and the accompanying translations show Norwegian verbs in similar constructions.

(1) He *seems disgruntled* about her reticence. (ABR1)
Han *virker irritert* over hennes tilbakeholdenhet.
(2) But none of my questioners *seems to have* any teeth. (FW1)
Men ingen av mine utspørrere *synes å ha* noen tenner.
(3) …more and more *it seemed that* two peoples lived in England, not one … (DL1)
…*det så mer ut til at* det bodde to folk i England, ikke ett.

Seem has been studied in contrast with Norwegian and Swedish by Johansson (2001) and Aijmer (2009). Both studies identify a wide range of translation correspondences, the most frequent of which are lexical verbs (although e.g. adverbs and modal particles also occur). The most frequent Norwegian lexical verb correspondences were those illustrated above: *virke, se ut* and *synes.* The previous studies noted differences between fiction and non-fiction in the ENPC and the English-Swedish Parallel Corpus as to the frequency of *seem* and to some extent its translation patterns.

Johansson (2001) and Aijmer (2009) studied Norwegian and Swedish expressions only to the extent that they occur as either translations or sources of *seem*. The present investigation widens the scope by giving more attention to the Norwegian 'seem-type' verbs identified in Johansson (2001), and studying their use in original Norwegian. The investigation also includes *appear,* due to the close similarity between *seem* and *appear* (for instance, they appear in each other's definitions in the *Macmillan Dictionary*). Furthermore, the register difference is explored in more depth. However, since the non-fiction part of the ENPC/ESPC cannot be considered a register, the fiction part of the ENPC will instead be compared to academic prose from the KIAP corpus. KIAP, whose acronym stems from the Norwegian name of the project Cultural Identity in Academic Prose (Fløttum et al. 2006), is a comparable corpus of published academic articles in English, Norwegian and French within the disciplines economics, linguistics and medicine. For the present purposes only English and Norwegian linguistics will be used. The research questions are as follows:

- What are the frequencies, patterns and meanings of *seem, appear* and their correspondences *virke, synes* and *se ut* in Norwegian and English fiction and academic prose?
('Pattern' applies to the syntactic patterns of the verb phrases (e.g. *seem* + *to*-infinitive, *seem* + adjective phrase, *seem* + *that*-clause or *as if*-clause) as well as their co-occurrence with a dummy subject and/or an experiencer phrase.)
- How do these patterns and their meanings compare across lexical items, languages and registers?

The present study uses the ENPC primarily as a comparable corpus of original fiction texts, but considers the translation correspondences as a source of information on parallelisms between the different expressions. The translation paradigm of *seem* established in Johansson (2001) is considered a viable *tertium comparationis* for my study even though the mutual correspondence (Altenberg 1999) of the expressions was not calculated.

The material was retrieved by searching for all inflectional forms of *seem, appear, virke, synes, se ut* in ENPC fiction and KIAP. Up to four words were permitted between the verb *se* and the particle *ut*. All the lexemes except *seem* have uses other than the relevant constructions, such as intransitive *appear* (='occur'), *virke* in the sense of 'work' *synes* in the sense of 'think', and *se ut* as a free combination of verb and preposition (e.g. *se ut av vinduet* = 'look out of the window'). These were removed following manual analysis of the concordance lines.

A preliminary analysis of the ENPC data reveals that the Norwegian lexemes are translated by *seem/appear* to different extents. For example, the Norwegian phrasal verb *se ut*, which was by far the most common correspondence of *seem* in Johansson (2001), has *look (like/as if*) as its most common translation in ENPC fiction. However, both *look* and *seem* combine sensory perception and cognition (Malá 2013). There are frequency differences between the registers in both languages: taken together, the 'seem-verbs' are more frequent in academic prose than in fiction. For English, the difference is mainly due to *appear*, which is seven times more frequent in academic prose. In Norwegian academic prose, *synes* and *se ut* are responsible for the difference. For *appear* and *synes* this can probably be linked to the more formal style of academic prose, while this is unlikely for *se ut*. The frequencies of *seem* and *virke* are similar across the registers. In both languages, the 'seem-verbs' occur more commonly with copular function in fiction while the catenative function is more common in academic prose. *Seem* and *appear* have rather similar profiles as regards copular vs. catenative function but the Norwegian verbs differ more. The final study will include considerations of the interplay of construction type with meaning (e.g. Usonienė & Šinkūnienė 2013). Johansson (2001) found that experiencer phrases with *seem*-type verbs are more common in Norwegian than in English. The present material gives the impression that experiencers are less common in academic prose than in fiction, although this analysis remains to be completed.

References

Aijmer, Karin. 2009. *Seem* and evidentiality. *Functions of Language* 16:1, 63-88.

Altenberg, B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In H. Hasselgård & S. Oksefjell (eds), *Out of Corpora. Studies in honour of Stig Johansson,* 249-268. Amsterdam: Rodopi.

Fløttum, Kjersti, Trine Dahl and Torodd Kinn. 2006. *Academic Voices*. Amsterdam: Benjamins.

Johansson, Stig. 2001. The English verb *seem* and its correspondences in Norwegian. What seems to be the problem? In K. Aijmer (ed.), *A Wealth of English. Studies in honour of Göran Kjellmer*, 221-245. Göteborg: Acta Universitatis Gothoburgensis.

Lampert, Günther. 2011. SEEM: evidential, epistemic, or what else? A study in cognitive semantics. *International Journal of Cognitive Linguistics* 2, (1): 1-24.

Macmillan Dictionary https://www.macmillandictionary.com/.

Malá, Markéta. 2013. Translation counterparts as markers of meaning: The case of copular verbs in a parallel English-Czech corpus. *Languages in Contrast* 13:2, 170-192.

Mortelmans, Tanja. 2017. Seem-type verbs in Dutch and German. *Lijken, schijnen & scheinen*. In Juana I. Marín-Arrese, Gerda Haßler and Marta Carretero (eds), *Evidentiality Revisited: Cognitive grammar, functional and discourse-pragmatic perspectives*, 123-148. Amsterdam/Philadelphia: Benjamins.

Usonienė, Aurelija & Šinkūnienė, Jolanta. 2013 A cross-linguistic look at the multifunctionality of the English verb *seem*. In J. I. Marín Arrese, M. Carretero, J. Arús Hita & J. van der Auwera (eds), *English Modality. Core, periphery and evidentiality*, 281–316. Berlin: Mouton de Gruyter.

Corpora

ENPC – English-Norwegian Parallel Corpus: https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/

KIAP – Cultural Identity in Academic Prose: http://www.uib.no/fremmedsprak/23107/kiap-korpuset

Verbs of perception in English/French and the explicitation of evidentiality in translation

Daniel Henkel
Université Paris 8 Vincennes-St. Denis, TransCrit
daniel.henkel@univ-paris8.fr

1. Introduction

This study is part of a series of analyses the aims of which are to identify major lexical and syntactic differences between English and French in translation compared with original English and French, to determine where these disparities occur, and to evaluate the degree of correspondence between target-texts and target-language norms. Verbs of perception such as 'see', 'hear' in English and 'voir', 'entendre' in French have often been a subject of theoretical inquiry from a contrastive perspective (Guillemin-Flescher (1981), Miller (2003), Chuquet (2004), Grezka (2006), Dufaye (2014), Bardière (2015) inter alia) but have not been investigated systematically and quantitatively in an empirical corpus-based study. The present analysis seeks to provide a comprehensive account of major differences in the use of 'hear', 'see', 'voir' and 'entendre' by comparing corpora of original and translated English and French following Johannson's (2007) “bidirectional” model.

2. Methods

Public domain literary works and their translations from the late 19th-early 20th centuries were collected from Project Gutenberg and Noslivres.net (a clearing-house of French-language public domain repositories) and retained on a 1-text-per-author/translator basis. A total of 35 works by 35 different authors in original English, 35 authors in original French and the same number of translations (4 subcorpora × 35 authors/translators = 140) were thus compiled into a 13-million word corpus consisting of four 3.3-3.5m-word subcorpora, which were tagged for POS and lemma in TreeTagger. To the best of the author's knowledge, this is the largest bidirectional NLP-annotated corpus in English and French currently available.

Each verb was inventoried individually and in connection with tense, aspect, voice and modal auxiliaries:
--present, preterit, present perfect, pluperfect, progressive forms in English,
--present, imperfect, passé simple, plus-que-parfait in French,
--all modals, 'can' and 'could' specifically in English,
--'pouvoir' in French,
--passive voice,

Other potentially important contextual variables were detected by comparing n-grams recurring in at least two-thirds of the texts in each corpus:
--first person subjects 'I'/'je'
--the pronoun 'on' in French
--'never'/'ever'/'jamais'
--'hear+of'
--'entendre+parler/dire'
--indefinite direct objects

Data were converted into frequencies per 10,000 words (f/10k) for each text, rather than the corpus as a whole, so as to observe the range of variation between individual authors' or translators' styles. The Wilcoxon-Mann-Whitney rank-sum test was used to determine whether the differences were statistically significant, after which their effect-size was assessed using Cohen's d. Finally, the degree of inter-

linguistic influence between source- and target-texts was evaluated using Spearman's correlation coefficient.

3. Results

3.1. Hear/Entendre

Disparities between authors and translators were found to be statistically significant for both 'hear' (p=0.035) and 'entendre' (p=0.028) with effect-sizes in the small to medium range (d=-0.59 for 'hear', d=-0.41 for 'entendre').

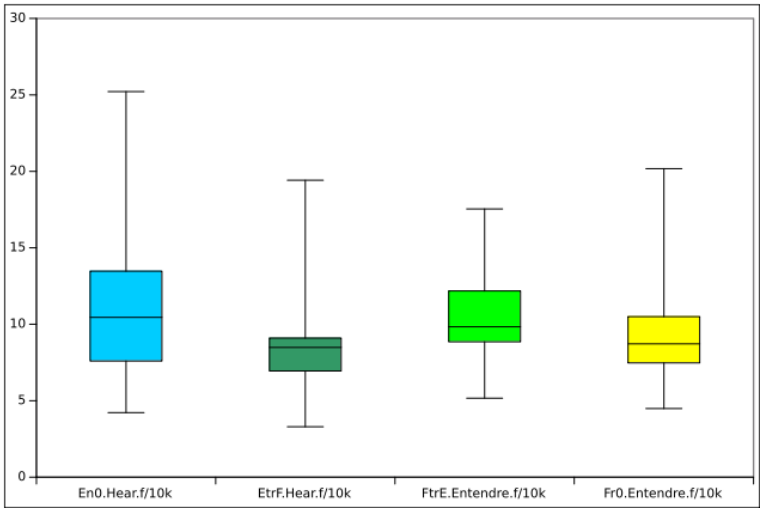


Fig. 1 *Range of frequencies (f/10k) for 'hear' and 'entendre': original English (En0), English-translated-from-French (EtrF), French-translated-from-English (FtrE), original French (Fr0)*

As shown in Figure 1, the interquartile range (IQR) for 'hear' in English-translated-from-French more closely resembles that of 'entendre' in original French with practically identical medians (f=8.72/10k for Fr0, f=8.48/10k for EtrF), while the median (f=9.84/10k) and IQR for 'entendre' in French-translated-from-English are closer to 'hear' in original English (median f=10.45/10k). Unsurprisingly, a strong correlation exists in both directions between source- and target-texts (rho=0.76 p<0.001 for FtrE/En0, rho=0.89 p<0.001 for EtrF/Fr0):

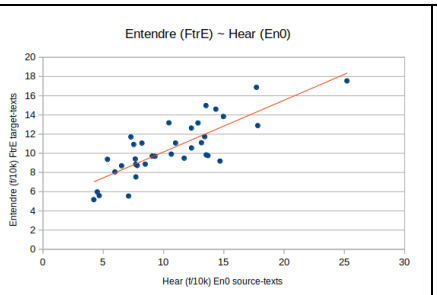


Fig. 2a *Frequency of 'hear' in En0 source-texts vs. frequency of 'entendre' in corresponding FtrE target-texts.*

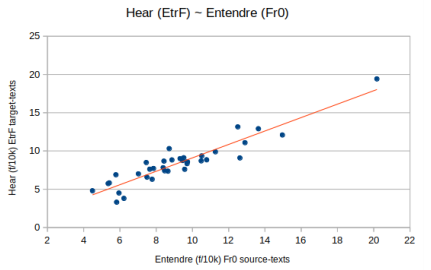


Fig. 2b *Frequency of 'entendre' in Fr0 source-texts vs. frequency of 'hear' in corresponding EtrF target-texts.*

Specific cases of over- or underuse, however, cancel each other out and are thus obscured in the overall frequencies. Although translators into English on the whole tend to use 'hear' less often than in En0, they actually use it more often in the passive voice 'be heard' (d=0.86, p<0.001). Conversely, while translators into French tend to use 'entendre' more often than in Fr0, they neglect to use it with 'on' (d=-0.76, p=0.002). The overuse of the passive in EtrF can, in fact, be attributed to a large extent to translations of “on

entendait…" etc. Moreover, the general tendency to use 'hear' less in EtrF and 'entendre' more often in FtrE is much stronger with certain expressions: 'hear of' (d=-1.19, p<0.001), 'entendre parler de' (d=1.21, p<0.001), 'entendre dire' (d=1.03, p<0.001).

3.2. See/Voir

Differences in the use of 'see' and 'voir' are more subtle. Although 'see' is used more frequently in English than in French, and the medians for translators are between those for authors, the range of variation and overlap between subcorpora is such that no significant difference in terms of overall frequency can be found between authors and translators:

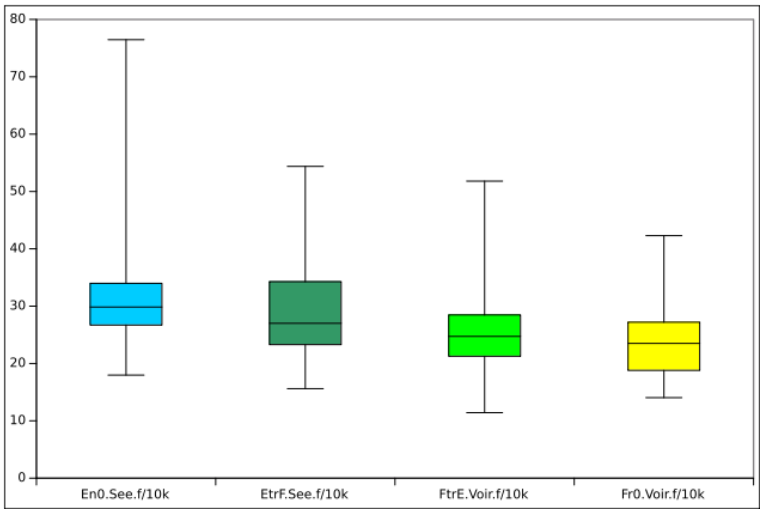


Fig. 3 *Range of frequencies (f/10k) for 'see' and 'voir': original English (En0), English-translated-from-French (EtrF), French-translated-from-English (FtrE), original French (Fr0)*

A relatively strong correlation (rho=0.63 p<0.001 for FtrE/En0, rho=0.8 p<0.001 for EtrF/Fr0) can again be seen between the frequencies of 'see' and 'voir' in corresponding source- and target-texts:

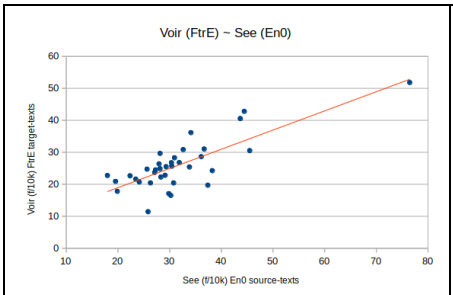


Fig. 4a *Frequency of 'see' in En0 source-texts vs. frequency of 'voir' in corresponding FtrE target-texts.*

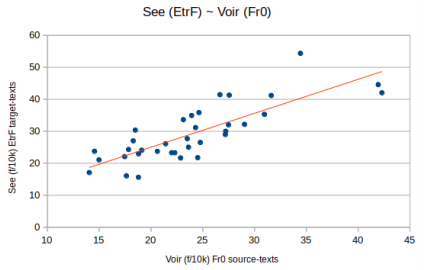


Fig. 4b *Frequency of 'voir' in Fr0 source-texts vs. frequency of 'see' in corresponding EtrF target-texts.*

Despite the lack of a clear difference in terms of overall frequency, significant disparities can nonetheless be observed between authors and translators in combination with certain variables. Again, the overuse of the passive form 'be seen' (d=0.66, p=0.001) runs counter to translators' overall tendency to use 'see' somewhat less often than in En0, while translators into French neglect to use 'voir' with 'on' (d=-0.76, p<0.001).

4. Discussion

Overall, verbs of perception are used more often in original English than in original French, while their frequency in translated texts is different enough that English-translated-from-French and French-translated-from-English can be recognized as distinct “subspecies”. The strong correlation coefficients suggest a high degree of interlinguistic influence or “shining-through” (Teich, 2003). These tendencies are more obvious with 'hear' and 'entendre', than with 'see' and 'voir', but in both cases the disparities are greater in combination with certain contextual variables, most noticeably the use of the passive voice in EtrF to translate 'on+voir/entendre'. More detailed analysis of specific examples, which could hardly be presented here for lack of space, will help to elucidate other circumstances in which verbs of perception may be overused in French-translated-from-English, or neglected in English-translated-from-French. The very concept of overuse or underuse, however, raises questions not only about the systematization of evidentiality in English and French, but about the deontology of explicitation/implicitation in translation as well.

References

Bardière, Y. (2015). Les traductions de can et could devant les verbes de perception. Modèles linguisitiques

Chuquet, J. (ed.) (2004). Verbes de parole, pensée, perception: Etudes syntaxiques et sémantiques, Rennes: PU de Rennes.

Dufaye, L. (2014). CAN avec les verbes de perception. in Autour du verbe anglais. Construction, lexique, évidentialité, Girard-Gillet, G. (éd.) Presses Sorbonne Nouvelle

Guillemin-Flescher, J. (1981). Syntaxe comparée de français et de l'anglais: problémes de traduction. Ophrys.

Grezka, A. (2006). Etudes du lexique de la perception: bilan et perspectives. Suvremena lingvistika, 61(1), 45-67.

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

McEnery, T., & 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/

Miller, P. & B. Lowrey, B. (2003). « La complémentation des verbes de perception en français et en anglais », In Miller, P., Zribi–Hertz, A. (eds.), Essais sur la grammaire comparée du français et de l’anglais, Paris: Presses Universitaires de Vincennes, pp. 133–188.

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

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

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

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Towards a better specification of the typological status of Italian in comparison with French
An analysis of its posture verbs

Thomas Hoelbeek
Vrije Universiteit Brussel
thomas.hoelbeek@vub.be

In this contribution I present a part of a research project that aims at better defining the typological position of Italian in comparison with French by tackling a previously unexplored aspect of these languages, viz. their posture verbs and expressions. Establishing how they function will help clarifying the typological status of both languages.

I focus in this paper on the classification of Italian within the 'verb-framed' vs. 'satellite-framed' language dichotomy (Talmy 1985). As a Romance language, Italian is expected to belong to the verb-framed class. However, it also shares important features with satellite-framed languages (cf Iacobini & Masini 2005; Iacobini & Fagard 2011) so that its exact status remains a matter of debate. The idea investigated is that the functioning of Italian posture verbs should be viewed as a satellite-framed phenomenon.

When applied to the domain of posture, the typological distinction between VF and SF languages is reflected in VF languages by a tendency to encode the location of entities by means of neutral verbs like *to be* (cf Hickmann 2007; Hickmann & Hendriks 2006; Lemmens & Slobin 2008), while the manner of posture tends to remain unexpressed. In contrast, SF languages are characterised by a tendency to specify the manner of posture by using specific verbs, e.g. in Dutch *staan* 'to stand', *liggen* 'to lie' or *zitten* 'to sit'. Also, it has been shown that, in SF languages, these verbs have developed complex semantic networks beyond the domain of static posture. This type of network has been observed for Dutch (Lemmens 2002; Lemmens & Perrez 2010), but also for German (cf De Knop & Perrez 2014), Swedish (Hellerstedt 2013) and, to a lesser extent, English (Newman 2002).

Interestingly enough, this phenomenon typically is not found in VF languages such as French. Although French does have posture predicates, their modern use seems mostly limited to encoding the posture of animate entities. However, this encoding is often dispensed with, since specifying the manner of posture can appear irrelevant (*Jean est (assis) sur le canapé* lit. 'John is (seated) on the couch').

The typological evolution from Satellite-framed Latin to Verb-framed Romance was a rather gradual transition though, characterised by different strategies and paces, which resulted in an important variation between Romance languages (Iacobini & Fagard 2011). Thus, for Italian we see that, while, as in French, specifying the manner of posture is rather irrelevant (*Giovanni è (seduto) sul divano* lit. 'John is (seated) on the couch'), a certain number of SF-like phenomena can be found as well. Some examples are the frequent use of directional post-verbal particles (satellites) associated with manner of motion verbs (cf Iacobini & Masini 2006; Cini 2008; Iacobini 2009), the size of the Manner of Motion lexicon (richer than those of the other VF languages (Iacobini 2010)), the presence of verbs that tend to merge Manner and Path (Lapesa & Lenci 2012) and the larger number of Manner of Speaking verbs as compared to, for example, English (Mastrofini 2014). Hence, all Romance languages are not equally VF, and Italian in particular manifests more SF characteristics than either French or Spanish (Iacobini & Fagard 2011).

The functioning of (static or dynamic) posture verbs in Italian (or in French for that matter) has not yet been studied though. In order to fill this gap a synchronic, semasiological analysis of the use of conventional static posture predicates is proposed (Italian *essere in piedi* lit. 'to be standing', *essere*

seduto lit. 'to be seated', *essere sdraiato/disteso* lit. 'to be lying' and their French equivalents *être debout*, *être assis*, *être couché*), which analyses also more marked static verbs such as French *seoir* 'to sit' which survives today in rather archaic expressions (*Derrière la maison sied un terrain de cinquante ares* 'Behind the house sits a fifty-acre plot of land').

Dynamic posture verbs and predicates are also included (Italian *mettersi in piedi* lit. 'to put oneself upright', *sedersi* lit. 'to sit oneself down', etc. and their French counterparts *se mettre debout*, *s'asseoir,* etc.). By relying on databases containing contemporary texts (for Italian, the *CORIS/CODIS* data base – containing 130 million words covering the 1980s and 1990s – and PAISÀ – composed of texts from the internet containing 250 million words collected in September/October 2010; for French, the reference data base *Frantext,* containing at present more than 5400 texts), a corpus study is conducted which will provide a precise description of the conditions of use and the semantic networks of the verbs at hand in written language.

To build up the corpus for this study, all the text fragments containing the aforementioned verbs are retrieved (for Italian the entire data bases are consulted, for French the interval 1980-2010 is selected in *Frantext*) and classified into semantic categories (the fragments consist of about 500 words, enlarged when necessary). Qualitative and quantitative analyses are combined, allowing to draw a detailed picture of the semantics involved, while providing objective results about frequency parameters. The qualitative angle offers a rich and detailed perspective, paying the same attention both to rare phenomena and more frequent ones. The tokens are considered one by one and specific semantic parameters are analysed, such as the Figure entity (animate – inanimate), the Ground entity, and the type of path (in case of dynamic configurations). In order to further compare the situation in the two Romance languages under study in an objective way, a more frequency-driven approach is relied on within the quantitative part of the study.

An early finding is that Latin *stare*, which lost its function of a posture verb in most Romance languages, is not found as a posture verb in contemporary French anymore (it only survived as forms of *être* 'to be'). Yet, in Italian it continues to exist as an auxiliary and copula, and is still used to describe the posture of both animate and inanimate entities (cf. *La macchina sta nel garage* 'The car stands in the garage'). This points to a more SF-like behaviour of Italian regarding the functioning of its posture verbs, which seems to confirm our starting hypothesis.

References

Cini, M. (ed.). 2008. *I verbi sintagmatici in italiano e nelle varietà dialettali* (Stato Dell'arte E Prospettive Di Ricerca Atti Delle Giornate Di Studio Torino, 19.-20. Febbraio 2007). Vol. 3. Bern: Peter Lang.

De Knop, S. & J. Perrez. 2014. Conceptual metaphors as a tool for the efficient teaching of Dutch and German posture verbs. *Review of Cognitive Linguistics* 12(1). 1–29.

Hellerstedt, M. 2013. *L'utilisation et l'acquisition des verbes de position en suédois L1 et L2.* (PhD Dissertation). Paris: Université de Paris-Sorbonne.

Hickmann, M. 2007. Static and dynamic location in French: Developmental and crosslinguistic perspectives. *The categorization of spatial entities in language and cognition* 20. 205–231.

Hickmann, M. & H. Hendriks. 2006. Static and dynamic location in French and in English. *First Language* 26(1). 103–135.

Iacobini, C. 2009. The role of dialects in the emergence of Italian phrasal verbs. *Morphology* 19(1). 15–44.

Iacobini, C. 2010. The number and use of manner verbs as a cue for typological change in the strategies of motion events encoding. *Space in language. Proceedings of the Pisa International Conference, ETS, Pisa* 495–514.

Iacobini, C. & B. Fagard. 2011. A diachronic approach to variation and change in the typology of motion event expression. A case study: From Latin to Romance. *Faits de langues: les cahiers* 3. 152–171.

Iacobini, C. & F. Masini. 2005. Verb-particle constructions and prefixed verbs in Italian: typology, diachrony and semantics. In *On-line Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5)*, 157–184.

Iacobini, C. & F. Masini. 2006. The emergence of verb-particle constructions in Italian: locative and actional meanings. *Morphology* 16(2). 155–188.

Lapesa, G. & A. Lenci. 2012. Italian Verbs of Manner of Motion at the Syntax-Semantics Interface: a Distributional Analysis. *Linguistic Evidence.* 9–11.

Lemmens, M. 2002. The semantic network of Dutch posture verbs. *Typological Studies in Language* 51. 103–140.
Lemmens, M. & J. Perrez. 2010. On the use of posture verbs by French-speaking learners of Dutch: A corpus-based study. *Cognitive Linguistics* 21(2). 315–347.
Lemmens, M. & D. I. Slobin. 2008. Positie-en bewegingswerkwoorden in het Nederlands, het Engels en het Frans. *Verslagen en mededelingen van de Koninklijke Academie voor Nederlandse Taal-en Letterkunde* 118(1). 17–32.
Mastrofini, R. 2014. English Manner of Speaking Verbs and their Italian Translations: A Cross-linguistic Comparison. *Athens Journal of Philology* 83–98.
Newman, J. 2002. A cross-linguistic overview of the posture verbs "sit", "stand" and "lie." In J. Newman (ed.), *The Linguistics of Sitting, Standing and Lying*, 1–24. Amsterdam-Philadelphia: John Benjamins.
Talmy, L. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description* 3. 57–149.

Register Analysis of Modal Verbs in Student Translations: A Corpus-based Study

Tiffany Jandrain
Université de Mons
tiffany.jandrain@umons.ac.be

While a tendency to examine registers can be increasingly observed in Translation Studies (to name a few, Baker, 1992/2011; Delaere, 2014; Gambier, 2013; Granger, 2016; Hansen-Schirra et al., 2012; Hatim & Mason, 1990; House, 1997/2015; Lefer & Vogeleer, 2016; Neumann, 2016; Schäffner, 2002; Steiner, 1998; Tomaszkiewicz, 2007; Trosborg, 1997b, 2002), many scholars are still calling for further investigation of the issue, as Johansson did in the early 2000s (Lefer & Vogeleer, 2016). In fact, the importance of analyzing registers before translating texts has recently been emphasized again (Gledhill & Kübler, 2016), since register analysis helps translators "gain increasing knowledge of socio-cultural variation in (specific) features pertaining to a genre in a particular culture" (Trosborg, 1997a, p. XV-XVI).

This paper tries to give this call some answers by presenting a corpus-based register study. It focuses on register analysis from the point of view of students translating from English into French. In fact, it has been claimed and proven that translation students have difficulty transposing register features and often make register mistakes in their translations partly because of the differences, in our case, between the English and French language systems (Chuquet & Paillard, 1987; Hewson, 1996). Indeed, whereas students often do feel that adapting language use according to the audience and the situation might be necessary to fulfil communicative functions of the text, many fail to do this successfully.

More specifically, the study of this paper is carried out on a specific linguistic feature that is relevant to register analysis in translation: modal verbs. In fact, since modality is within the scope of the enunciation approach and therefore gives information on how objects are referred to in situational context (Branca-Rosoff, 1999), modality analysis is thought to help students consider components of register when they translate. More precisely, it gives students a view of how necessity and possibility moods are used in these registers in French, and, more precisely, how their different forms, according to the "traditional classification" (alethic, deontic and epistemic modalities) (Saussure, 2014), are used.

In Translation Studies, Halliday's model of discourse analysis from the perspective of Systemic Functional Linguistics was employed by several scholars in their works (to name a few, Baker, 1992/2011; Hatim & Mason, 1990; House, 1997/2015; Munday, 1998; Neumann, 2016; Steiner, 1998; Trosborg, 2002). The theory of Functionalism in translation has considered the issue of register variation as well, with Nord (2005)'s translation text-oriented analysis model in particular. Nevertheless, while Register analysis approaches focus on a discourse analysis and Functional theories of translation on a textual analysis, it is suggested here that translation students may receive benefit from a combination of both approaches in their register analysis. We follow Biber and Conrad (2009)'s definition of register, which is "a language variety associated with both a particular situation of use and with pervasive linguistic features that serve important functions within that situation of use" (p. 31). Since the interpretation of modal features directly depends on the situational context (Saussure, 2014), Biber and Conrad's framework of register analysis appears appropriate for the study of modality.

Nevertheless, it was necessary to adapt this framework for our comparative and translation purposes, as it was specifically established for the analysis of English registers: 1) its linguistic categories of analysis were modified for the consideration of French texts (Riegel et al.'s *Grammaire méthodique du français* (2018) was chosen for this purpose); 2) Biber and Conrad's step of interpreting functionally linguistic features of a register with reference to its situational context (which they hardly explain in their framework)

was studied in this paper on the enunciative approach, which is not formally considered in the framework; 3) a translation perspective was added as well.

Besides, since it can be hypothesized that translated texts can present linguistic differences from non-translated texts (Zanettin, 2013, in Kruger, 2018), this study also implied a study on corpora of texts which were considered comparable to those that the students were asked to translate.

In other words, the data for this study has been retrieved from four French corpora, which were compiled for this study: two corpora of 14 translations (of a press article and a popular science article) made by advanced students into their L1, and two corpora of 30 non-translated texts (press articles and popular science articles) written by experts in their L1. Modal verbs were identified in our corpora by using the Sketch Engine concordancer (Kilgarriff et al., 2014). Occurrences were annotated and verified by running an intercoder agreement test according to Spooren and Degand (2010)'s procedure (the results of Cohen's kappa (= 0.757 for *devoir* and = 0.635 for *pouvoir*) can be interpreted as "moderate" (McHugh, 2012))

Results of the comparable corpora show that modality is used differently in those registers (p<0.05). Regarding possibility, press articles tend not to imply addressees' alethic and deontic capacity, whereas popular science articles tend to include them along with addressors' capacity. Regarding necessity, alethic obligation is rarely used in both registers, whereas deontic obligation is overrepresented in press articles (p<0.05) and is used to animate abstract concepts in popular science articles.

Results of the translation corpora show that there is a significant difference in the use of deontic obligation by students in their translations of the press article (p<0.05), while there is no significant difference of all the other types of modality in their translations. Furthermore, 57.1% of the students tend to translate the verb *have to* into *devoir*, whereas the verb *can* is more often translated into *pouvoir* (by 71.4% to 100% of the students). It should also be noted that those two verbs are used in students' translations in the same way as they are used in the comparable corpora. Finally, Delizée (2012)'s typology is used to describe errors that students make. It appears that students who decided to reformulate the modal verbs of the source text tend to render incorrect meaning (*glissement de sens*) (Delisle et al., 1999). Concluding remarks and suggestions will be then presented.

References

Baker, M. (2011). In Other Words: A Coursebook on Translation (2e edition). Routledge.

Biber, D., & Conrad, S. (2009). Register, Genre, and Style. Cambridge University Press.

Branca-Rosoff, S. (1999). Types, modes et genres: Entre langue et discours. Langage et société, 87(1), 5–24. https://doi.org/10.3406/lsoc.1999.2851

Chuquet, H., & Paillard, M. (1987). Approche linguistique des problèmes de traduction. Ophrys.

Delaere, I. (2014). A corpus-based multivariate study into register differences between translated and non-translated Belgian Dutch [Doctoral dissertation]. UGent.

Delisle, J., Lee-Jahnke, H., & Cormier, M. C. (Eds.). (1999). Terminologie de la traduction—Translation Terminology—Terminología de la Traducción—Terminologie der Übersetzung. John Benjamins Publishing Company.

Delizée, A. (2012). L'évaluation formative en traduction: Que désire-t-on évaluer et comment ? Journée d'étude 'Traduction et qualité - Méthodologies en matière d'assurance qualité', Université de Lille III.

Gambier, Y. (2013). Genres, text-types and translation. In Y. Gambier & L. Van Doorslaer (Eds.), Handbook of Translation Studies (Vol. 4, pp. 63–69). John Benjamins Publishing Company.

Gledhill, C., & Kübler, N. (2016). What can linguistic approaches bring to English for Specific Purposes? ASp, Concepts and Frameworks in English for Specific Purposes(69), 65–95.

Granger, S. (2016). A lexical bundle approach to comparing languages. Stems in English and in French. In M.-A. Lefer & S. Vogeleer (Eds.), Genre- and Register-related Discourse Features in Contrast (pp. 59–72). John Benjamins Publishing Company.

Hansen-Schirra, S., Neumann, S., & Steiner, E. (2012). Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English-German. De Gruyter Mouton.

Hatim, B., & Mason, I. (1990). Discourse and the Translator. Longman.
Hewson, L. (1996). Le niveau de langue repère. Palimpsestes - Niveaux de Langue et Registres de La Traduction, 77–92.
House, J. (2015). Translation Quality Assessment. Past and Present. Routledge.
Kassambara, A. (2018). Kappa de Cohen dans R: Pour deux variables catégorielles. Data Novia : Mesures de la Concordance Inter-Évaluateurs dans R. https://www.datanovia.com/en/fr/lessons/kappa-de-cohen-dans-r-pour-deux-variables-categorielles/
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
Kruger, H. (2018). Expanding the third code: Corpus-based studies of constrained communication and language mediation. Using Corpora in Contrastive and Translation Studies (5th Edition), Université Catholique de Louvain.
Lefer, M.-A., & Vogeleer, S. (2016). Introduction. In M.-A. Lefer & S. Vogeleer (Eds.), Genre- and Register-related Discourse Features in Contrast (pp. 1–6). John Benjamins Publishing Company.
Munday, J. (1998). A Computer-assisted Approach to the Analysis of Translation Shifts. Meta, 43(4), 542–556.
Neumann, S. (2016). Cross-linguistic register studies. In M.-A. Lefer & S. Vogeleer (Eds.), Genre- and Register-related Discourse Features in Contrast (pp. 35–57). John Benjamins Publishing Company.
Nord, C. (2005). Text Analysis in Translation. Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis (2e edition). Rodopi.
Riegel, M., Pellat, J.-C., & Rioul, R. (2018). Grammaire méthodique du français (7e édition). Presses Universitaires de Paris (puf).
Saussure, L. (2014). Verbes modaux et enrichissement pragmatiques. Langages, 1(193), 113–126.
Schäffner, C. (2002). Discourse analysis for translation and translator training: Status, needs, methods. In C. Schäffner (Ed.), The Role of Discourse Analysis for Translation and in Translator Training (pp. 1–8). Multilingual Matters LTD.
Spooren, W., & Degand, L. (2010). Coding coherence relations: Reliability and validity. Corpus Linguistics and Linguistic Theory, 6(2). https://doi.org/10.1515/cllt.2010.009
Steiner, E. (1998). A Register-based Translation Evaluation: An Advertisement as a Case in Point. Target, 10(2), 291–318.
Tomaszkiewicz, T. (2007). Transfert de différents registres de la langue parlée. In C. Wecksteen & A. El Kaladi (Eds.), La traduction dans tous ses états (pp. 161–176). Artois Presses Université.
Trosborg, A. (1997a). Introduction. In A. Trosborg (Ed.), Text Typology and Translation (pp. vii–xvi). John Benjamins Publishing Company.
Trosborg, A. (1997b). Text Typology: Register, Genre and Text Type. In A. Trosborg (Ed.), Text Typology and Translation (pp. 3–23). John Benjamins Publishing Company.
Trosborg, A. (2002). Discourse Analysis as Part of Translator Training. In C. Schäffner (Ed.), The Role of Discourse Analysis for Translation and in Translator Training (pp. 9–52). Multilingual Matters LTD.

On Similative Demonstratives in Czech and English: Evidence from Corpora

Markéta Janebová, Michaela Martinková
Palacký University Olomouc
marketa.janebova@upol.cz, michaela.martinkova@upol.cz

Assessment of similarity is regarded as one of the basic cognitive activities (e.g. König 2017). This paper is a contrastive corpus-based study of one of the means of expressing similarity, i.e. demonstratives which express similarity (similative demonstratives, SDs). It explores "atypical uses" of the Czech demonstrative *takový*; the question we pose is whether these atypical uses correspond to those of its English dictionary equivalent *such.* The study uses three methods: analysis of (i) comparable corpora, (ii) parallel corpora, and (iii) monolingual corpora.

Demonstratives are a typologically well-established grammatical category (e.g. Diessel 2006) with a range of functions which often go "beyond their well-described exophoric and endophoric ones" (König 2017); these are called "atypical" uses. The sources of these uses can be found in the basic semantic structure of demonstratives, which includes the categories of manner, quality, and degree (MQD) and allows demonstratives to express similarity (hence similative demonstratives) and to create ad-hoc kinds (König and Umbach 2017).

So far, paths of grammaticalization have been proposed for demonstratives of manner and degree (cf. König 2017); this paper focuses on quality, expressed by *such* and *takový.* The phoric use of *takový*, a derived demonstrative pronoun with adjectival inflection (cf. Komárek et al. 1986), is exemplified in (1), where it expresses a contextually retrievable quality:

(1) *Takových* lidí je mnoho.
of-SD people are many
'There are many such people.'

The Dictionary of Standard Czech (*SSJČ*) also lists several non-phoric uses of *takový*: e.g. degree (2), indefinite use (according to *SSJČ*, it expresses "embarrassment stemming from the inability to provide a more specific description"; e.g. (3)), and a definite but unspecified description (4):

(2) Mám *takový* hlad!
I-have SD hunger
'I am so hungry!'

(3) Je *takový* divný.
he-is SD strange
'He is kind of strange.'

(4) Přijeli na *takový* zámek o samotě ve velké zahradě.
they-came to SD castle in isolation in big garden
'They came to this isolated castle in a big garden.'

(4) exemplifies the "recognitional" use: it signals to the hearer that "the speaker is referring to specific, but presumably shared, knowledge" (Himmelmann 1996:240).

From the non-phoric uses of *takový*, only the intensifying use (i.e. degree) is posited for *such* by van der Auwera and Coussé (2016). The multitude of the meanings of *takový* raises the question of whether they are all autonomous, and whether they all based on similarity like the English *such* (which has "a variety of uses, which nearly always involve the expression of similarity" [ibid. 15]).

(i) As a first step, we created comparable corpora of original Czech and English fiction (997,103 and 1,096,663 tokens, respectively) in order to establish the degree of mutual correspondence (MC) of *such* and *takový*. *Takový* is more frequent than *such* (1,090 tokens of *takový* vs. 430 tokens of *such*); the difference is statistically significant ($\chi 2 = 352.87$, $p < 0.001$). The MC is 30.3%, with a clear translation bias (48.8% cases of *such* are translated by *takový*, but only 23% cases of *takový* are translated by *such*).

(ii) In the next step, we examined the discourse functions of *takový* through its English translations in a subcorpus of original Czech post-1950 fiction, created on the basis of InterCorp 10 (Klégr et al. 2017; 2,960,365 text positions, sentence-aligned with the published English translations). There were 2,663 tokens of *takový* (899.66 pmw); a random sample of 500 tokens was subjected to qualitative and quantitative analyses.

Each token was coded for its syntactic structure (e.g. the presence of a demonstrative pronoun, noun, adjective), context (anaphoric reference), English translation equivalent, and the discourse function of *takový* (e.g. quality, degree, recognitional). At this stage of research, we subsumed the "embarrassment" indefinite meaning exemplified in (3) under the recognitional use (a very expanded context would be necessary for further coding).

Quality represents more than 50% in the sample, and degree about 20%. The recognitional use represents more than 20%, while the other non-typical uses are marginal.

In order to get a more precise idea of the role played by the variables, we fitted a logistic regression model, which shows that another demonstrative and/or an adjective in the phrase are significant predictors for the recognitional use. As for translation equivalents of *takový*, we found significant overrepresentation of the indefinite article and type nouns such as *kind* or *sort* with the recognitional use. On the other hand, *such* was significantly underrepresented with the recognitional use and overrepresented with the quality use.

(iii) Finally, we used the monolingual corpus of spoken Czech (Oral v.1) to see how these discourse functions are represented in spontaneous dialogues. *Takový* is almost five times more frequent in spoken Czech than in fiction (4,442.68 pmw). The quality use is less frequent (14%), while the recognitional use dominates (79%).

In our spoken data speakers also use *takový* when unable to retrieve a lexeme for a lemma in their mental lexicon, or hesitate to use a lexeme since they are not sure whether it denotes the lemma retrieved. It is often followed by pauses, hesitation markers or hedges such as *jako* 'like'. According to Consten and Averintseva-Klisch (2012:262), it is sometimes "not clear whether it is really the reference that matters or the correct lexical choice".

These results suggest that the recognitional use is an autonomous and not really "atypical" function in Czech. It is associated with English type nouns as translation equivalents, which tend to develop into markers of imprecision, i.e. hedges: they acquire "pragmatic, interpersonal, and speaker-based functions" (Traugott 1995:32). In this respect, English type nouns as a translation equivalent of *takový* attest to its hedging and interpersonal functions which *such* does not have (hence the translation bias).

Spoken data call for a neater discrimination within the use coded as recognitional, and for more quantitative analyses of correlations between the functions and their formal realizations. This suggests that we will need to look into the difference between the indefinite and recognitional uses of

demonstratives in greater detail. Ultimately, an analysis of diachronic data will be needed to confirm or reject the proposed paths of change.

References

van der Auwera, J., & Coussé, E. (2016). *Such* and *Sådan* – the Same but Different. *Nordic Journal of English Studies*, 15(3), 15–32.

Consten, M., & Averintseva-Klisch, M. (2012). Tentative Reference Acts? "Recognitional Demonstratives" as Means of Suggesting Mutual Knowledge – or Overriding a Lack of It. *Research in Language*, (10)3, 257–77.

Diessel, H. (2006). Demonstratives, Joint Attention, and the Emergence of Grammar. *Cognitive Linguistics*, 17, 463–89.

Himmelmann, N. (1996). Demonstratives in Narrative Discourse: A Taxonomy of Universal Uses. In B. Fox (Ed.), *Studies in anaphora* (pp. 205–54). Amsterdam: Benjamins.

Hirschová, M. (1988). Netypické případy užití ukazovacích výrazů *takový*, *tak* [Atypical uses of the demonstratives *takový*, *tak*]. *Naše řeč*, 71(2), 57–61.

Klégr, A., Kubánek, M., Malá, M., Rohrauer, L., Šaldová, P., & Martin Vavřín. (2017). *Korpus InterCorp – angličtina, verze 10 z 1. 12. 2017*. Ústav Českého národního korpusu FF UK, Praha. Available at http://www.korpus.cz.

Komárek, M., Kořenský, J., Petr J., Veselková, J., et al. (1986). *Mluvnice češtiny 2. Tvarosloví* [Grammar of Czech 2. Morphology]. Prague: Academia.

König, E. (2017). The Deictic Identification of Similarity. In Y. Treis & M. Vanhove (Eds.), *Similative and Equative Constructions: A Cross-linguistic Perspective* (pp. 143–164). Amsterdam: Benjamins.

König, E., & Umbach, C. (2017). Demonstratives of Manner, of Quality and of Degree: A Neglected Subclass." In M. Coniglio, A. Murphy, E. Schlachter & T. Veenstra (Eds.), *Atypical Demonstratives: Syntax, Semantics and Pragmatics*. Berlin: de Gruyter Mouton. Final draft. Available at http://www.carla-umbach.de/publications/Koenig_Umbach_revised.Nov2015.pdf

Slovník spisovného jazyka českého [SSJČ; Dictionary of Standard Czech]. (1989). Ed. B. Havránek. Praha: Academia.

Traugott, E. C. (1995). Subjectification in Grammaticalization. In D. Stein & S. Wright (Eds.), *Subjectivity and Subjectivisation: Linguistic Perspectives* (pp. 31–54). Cambridge: Cambridge University Press.

"I'm not sure, how can I say?" A cross-cultural study of expressions of uncertainty by Italian and Norwegian learners and native speakers of English

Stine Hulleberg Johansen[1], Francesca Poli[2]
University of Oslo[1], Università Cattolica del Sacro Cuore[2]
s.h.johansen@ilos.uio.no, francesca.poli@unicatt.it

Introduction

Communicating interactively is a challenging task for L2 learners who are required to produce appropriate responses in rapidly developing discourse (Gablasova et al., 2017). This may lead them to feel uncertain and reflect that uncertainty in speech. However, how this uncertainty is conveyed and the degree to which it is expressed may vary across different cultural groups. In Norway, for instance, Gray (2005) claimed that it is acceptable to admit a lack of knowledge about a given subject. On the contrary, Italy is characterised by high Uncertainty Avoidance resulting in the need for clear and interpretable language (Hofstede, 2010). This study compares the use of expressions of uncertainty (ExU) in spoken English between two culturally different (Hofstede, 2010) groups of L2 learners, Italian and Norwegian, and a group of native English speakers. Our purpose is to explore whether non-native speakers' ExU differ from those of native speakers and across different non-native groups, following Granger's Contrastive Interlanguage Analysis (2015). The study contrasts two L2 and one L1 variety of English by utilising recently compiled and previously unexploited spoken learner corpus data and applying a partly corpus-driven and corpus-based approach. The work addresses the following research question:

Are there any differences in the frequency or type of ExU produced by Italian learners, Norwegian learners and native speakers of English?

Rationale

When we speak, we not only communicate propositions, but also attitudes to these propositions. Expressing uncertainty is closely linked to epistemic modality, i.e. the speaker's judgements or assumptions about the factual status of a proposition (Coates, 1987), and the concept of stance, i.e. the expression of "attitudes, thoughts and feelings of the speaker" (Biber et al., 1999: 966).

Studying ExU in learners' speech is interesting as extensive pragmatic competence is required to appropriately express one's commitment to an assertion (Holmes, 1982), and studies have shown that even advanced learners tend to exhibit a limited repertoire of pragmatic resources (Romero-Trillo, 2018). Other research has suggested differences between learners and native speakers in the use of adverbs of certainty (Perez-Paredes & Camino Bueno-Alastuey, 2019). Nevertheless, the pragmatics of spoken communication in general remains under-researched (Gablasova et al., 2017). This is particularly true for corpus-based studies of Italian and Norwegian learners, due to the limited number of spoken corpora available.

Methodology

In line with the need for more corpus-driven research (Callies, 2015), *n*-grams were first extracted from three comparable corpora: the Italian Spoken Learner Corpus (ISLC) (Poli, 2020), the Norwegian component of LINDSEI (Aas & Nacey, 2019), and the native-speaker reference corpus LOCNEC (De Cock, 2004). The ISLC and LINDSEI-no contain data from ≥ C1 learners of English. The minimum *n*-gram size for the extraction was set to min. two and max. five; the frequency threshold was set to five occurrences and a minimum distribution of three speakers. This initial search resulted in an inventory of approximately 20,000 *n*-grams overall. These were

manually sorted and cleaned from any irrelevant occurrences (e.g. *but I*, *overlap and*) resulting in 12 simplified expressions: *I think*, *I don't think*, *I'm not sure*, *I don't know*, *I would say*, *I guess*, *I suppose*, *maybe*, *probably*, *perhaps*, *let's say*, *how can I say*. Then, their frequency was checked and irrelevant instances were discarded. The final dataset included the speakers divided by L1 group with their individual relative frequency per 100,000 words for each of the 12 expressions. To address our research question, Kruskal-Wallis tests followed by post-hoc pairwise Wilcoxon rank sum tests with Holm correction were carried out in R.

Findings

The results depict a complex picture of the use of uncertainty: aside from *I think*, *I don't think*, *I don't know* and *I would say* which are not statistically different across the three groups, there was statistical difference between groups in *I'm not sure* ($H(2) = 7.41$, $p = 0.025$), *I guess* ($H(2) = 22.66$, $p = < 0.0001$), *I suppose* ($H(2) = 54.81$, $p = < 0.0001$), *maybe* ($H(2) = 53.39$, $p = < 0.0001$), *probably* ($H(2) = 11.58$, $p = 0.003$), *perhaps* ($H(2) = 17.45$, $p = < 0.0001$), *let's say* ($H(2) = 21.55$, $p = < 0.0001$), and *how can I say* ($H(2) = 8.56$, $p = 0.014$). The post-hoc tests indicated that the Norwegian learners overuse *I'm not sure*, *I guess*, and *maybe*, while the Italian speakers overuse *probably, I guess, maybe, perhaps, let's say, how can I say*. Both groups of learners overuse *maybe* and *I guess*, while they seem to underuse *I suppose* which stands as a typically British English expression. The Norwegian learners display a more similar behaviour to the native speakers with less quantitative differences compared to their Italian peers.

Conclusion

In conclusion, the Italian learners display a higher degree of uncertainty that sets them apart from their Norwegian and British peers. The results show overlap for a number of expressions which are produced in similar amounts by all groups, while Italian learners overuse five ExU. Both Italians and Norwegians underuse *I suppose*. On the other hand, the Norwegian learners make a similar use of ExU to the native English speakers, except for the overuse of *I guess*, *I'm not sure,* and *maybe*.

Although additional (L1 contrastive) research is needed to better frame the complex pattern highlighted by this analysis, it could be hypothesised that Italians are generally more unsure in English compared to Norwegian and native speakers, despite advanced proficiency and their high level of Uncertainty Avoidance (Hofstede, 2010). Their overuse of *let's say, how can I say* could also be an instance of interpersonal conditional transferred to the L2. The overuse of *I guess* and *I'm not sure* by Norwegians could be a result of L1 transfer as there are corresponding expressions in Norwegian, but closer scrutiny is needed to rule out other explanations such as pragmatic fossilization (Romero-Trillo, 2018) or personalization of talk (Baumgarten & House, 2010).

References

Aas, H. L., & Nacey, S. (2019). Methodological Concerns for Investigating Pause Behavior in Spoken Corpora. In L. Degand, G. Gilquin, L. Meurant, & A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties* (pp. 41–64). Louvain-La-Neuve: Presses universitaires de Louvain.

Baumgarten, N. & House, J. 2010. *I think* and *I don't know* in English as lingua franca and native English discourse. *Journal of Pragmatics. 42*. 1184–1200.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. & Quirk, R. (1999). *Longman grammar of spoken and written English*. Longman.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, (pp. 35–55). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003

Coates, J. (1987). Epistemic Modality and Spoken Discourse. *Transactions of the Philological Society*, 85, 110–131. https://doi.org/10.1111/j.1467-968X.1987.tb00714.x

De Cock, S. (2004). Preferred Sequences of Words in NS and NNS Speech. *Belgian Journal of English Language and Literatures* (BELL) (New Series 2), 225–246.

Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (2017). Epistemic Stance in Spoken L2 English: The Effect of Task and Speaker Style, *Applied Linguistics*, *38*(5). 613–637. https://doi.org/10.1093/applin/amv055

Gilquin, G., De Cock, S. & Granger, S. (2010). *LINDSEI. Louvain International Database of Spoken English Interlanguage*. Presses universitaires de Louvain.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24.

Gray, S. (2005). *Kalde nordmenn? Norske høflighetsnormer krysskulturelt perspektiv* [*Cold Norwegians? Norwegian politeness norms in a cross-cultural perspective*], Master thesis. Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo.

Hofstede G., Hofstede G. J. & Minkov M. (2010). Cultures and organizations: Software of the mind, intercultural cooperation and its importance for survival, 3rd edition, McGraw-Hill, New York.

Holmes, J. (1982). Expressing Doubt and Certainty in English. *RELC Journal*, 13(2), 9–28. https://doi.org/10.1177/003368828201300202

Perez-Paredes, P. & Camino Bueno-Alastuey, M. (2019). A corpus-driven analysis of certainty stance adverbs: *Obviously*, *really* and *actually* in spoken native and learner English. *Journal of Pragmatics* 140, 22–32. https://doi.org/10.1016/j.pragma.2018.11.016

Poli, F. (2020). Adverb + adjective collocations in a spoken learner corpus: A quantitative and qualitative approach [unpublished doctoral dissertation]. Università Cattolica del Sacro Cuore.

Romero-Trillo, J. (2018). Corpus Pragmatics and Second Language Pragmatics: A Mutualistic Entente in Theory and Practice. *International Journal of Corpus Linguistics and Pragmatics*, *2*(2), 113–127. https://doi.org/10.1007/s41701-018-0031-5

Formulaicity in constrained texts: an intermodal approach

Marta Kajzer-Wietrzny[1], Łukasz Grabowski[2]
Adam Mickiewicz University[1], University of Opole[2]
kajzer@amu.edu.pl, lukasz@uni.opole.pl

The quest for translation universals has recently turned to identification of specific patterns in other forms of language contact or constrained communication. For example, Lanstyák and Heltai (2012) hypothesize that both translation and non-native production share the main constraint, i.e. the need to manage two languages and the ensuing "linguistic uncertainty resulting from the parallel activation of two languages". At the same time, they point out that constrained varieties differ in that non-native language/text production involves descriptive language use (i.e. it does not depend on any other text), translation is additionally constrained by interpretive language use (i.e. it is dependent on the source text).

Studies of translated English and non-native English point to similar linguistic tendencies with respect to "features resulting from processing strain" (Kruger & Van Rooy 2016a: 26). Among the constrained varieties, translation is usually viewed as the extreme case of bilingual activation and perceived as particularly constrained at the psycholinguistic level due to rapid bi-directional switching between languages and activation both at the level of language in general as well as the specific linguistic variants of the source text (Kruger & van Rooy 2016b: 121). From this perspective, simultaneous interpreting is an even more extreme case due to the time constraint, which makes the entire process more rapid than in written translation. It is thus vital to expand the analysis to interpreting as in many respects it shows different linguistic patterns than translation (cf. Shlesinger & Ordan 2012; Defrancq et al. 2015; Ferraresi et al. 2019). Likewise, spoken non-native texts should be included in this paradigm, because like interpreting, such texts are not subject to editing.

Literature on constrained communication points to shared cognitive limitations in the production of non-native and translated texts and, as pointed by Aston (2018 84-85, after Forster 2001), cognitive resources seem to be liberated by the use of formulae which are also believed to be used in greater proportions in settings requiring more processing effort. In an exploratory study of interpreter discourse at the European Parliament, Aston (2018: 83) looks at the frequency of n-grams with 5 words or longer found in transcripts of simultaneous interpretations and argues that "the language of fluent interpreters relies heavily on recurrent formulaic phraseologies." As the formulaic repertoire of second language speakers is supposed to be smaller than that of native speakers, Aston (2018:83) points "to the need for interpreters working into their second language to enlarge this repertoire as far as possible", especially that linguistic preferences of translators and interpreters do not always reflect native speakers' preferences.

In view of this observation, in this study, positioned on the interface between corpus linguistics, formulaic language and Translation/Interpreting Studies we look at the formulaicity of texts produced by native English speakers and native speakers of Polish using English as well as Polish interpreters at the European Parliament working into their B (L2) language and Polish-English translations of the European Parliament debates. The study aims to verify whether constrained texts differ from native texts in terms of the use of adjacent word combinations commonly known as bigrams and whether similar patterns can be found across spoken and written registers.

The research material includes the Polish-English components of the European Parliament Translation and Interpreting Corpus (https://corpora.dipintra.it/eptic/), which is an intermodal corpus rich in contextual information (e.g. speaker, delivery rate, mode of delivery of the text/source text, nativeness).

Spoken	Written				
Native English Originals*	Non-native English Originals**	Interpretations from Polish into English*	Native English Originals*	Non-native English Originals**	Translations from Polish into English*
9,487 w 34 texts	9,869 w 33 texts	9,567 w 58 texts	9,200 w 34 texts	9,703 w 33 texts	11,714 w 58 texts

* *Components of EPTIC*
** *Corpora compiled according to EPTIC guidelines*

Figure 1. Data description (cf. Kajzer-Wietrzny & Grabowski 2021:158).

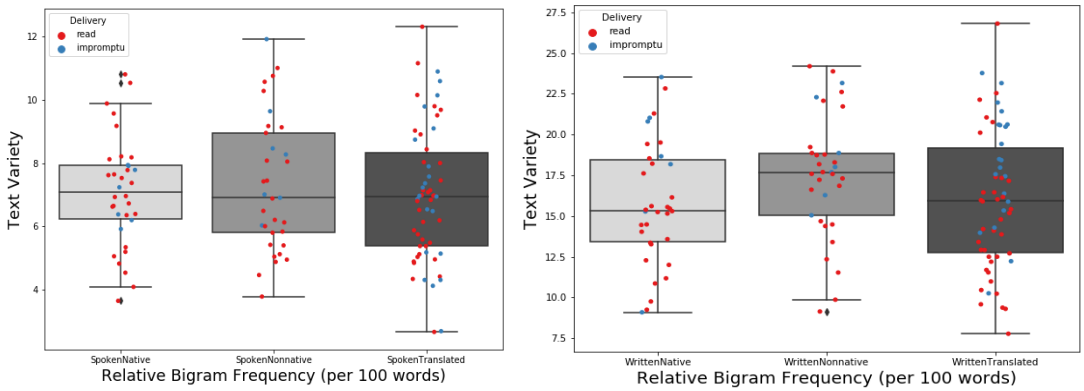


Figure 2. Frequency of bigrams in analysed datasets with respect to text variety and mode of delivery.

We resort to bigrams as the unit of analysis because they have been effectively applied in modelling language data in various statistical NLP tasks, and as indicators of formulaic language in texts (Altenberg 1998). Bigrams tap into the most important aspects of formulaic language (from the corpus linguistic perspective seen primarily as recurrent use of fixed or semi-fixed multi-word units in texts), such as frequency and fixedness (Schmitt & Carter 2004, Wood 2015, Pęzik 2018, Siyanova-Chanturia & Omidian 2019). Also, the frequency-driven approach to study formulaic language is particularly useful for the analyses of clichéd texts because such texts rely more on limited stocks of prefabricated text chunks or boilerplate conventional formulas (Forsyth & Grabowski 2015). Also, Nesi (2012: 422) claims that "n-grams in spoken and written texts tend to be constituted differently [...], and some genres are more formulaic than others".

To address this issue we fitted Poisson regression models, applicable to count variables (Winter 2019), with fixed and random effects in R (2013) using lme4 package. The total count of the most frequent bigrams was modelled as a function of predictor variables: text variety and mode of delivery (and delivery rate in the spoken subset) of the source adjusted by an exposure variable, i.e. total number of bigrams in text. Text-specific random intercepts were also included for the effect of text variety, mode of delivery and delivery rate (in the spoken subset) on the count of the most frequent n-grams. We hypothesize that due to increased processing constraints interpreters, translators and non-native speakers rely more on the use of formulaic expressions (here bigrams) than native speakers, and that the mode of delivery of the text and delivery rate (in the case of spoken production) might impact the bigrams' frequency. The tools used in the study include Formulib software package (Forsyth 2015), R (2013) and ad hoc scripts written in Python.

Our results show that the translated language variety indeed contributes to the increased use of the most frequent bigrams in spoken ($p=0.00584$) and written ($p=0.02680$) registers. A similar trend, albeit not a significant one, can be observed in the non-native production in both registers. Moreover, the number of frequent bigrams in texts generally increases when the speech/source speech is delivered impromptu,

but the effect is significant only for the written register (p=0.00291). The results also reveal considerable impact of individual variation on formulaicity as most of the variation within both models is explained by the text-specific random variables rather than the fixed variables.

References

Altenberg, B. (1998). "On the phraseology of spoken English: The evidence of recurrent word combinations". In: A. Cowie (Ed.), Phraseology: Theory, Analysis and Applications. Oxford: Oxford University Press, 101–122.

Aston, G. (2018). "Acquiring the language of interpreters: A Corpus-based Approach." Making Way in Corpus-based Interpreting Studies. Singapore: Springer, 83-96.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2014). "Fitting linear mixed-effects models using lme4". arXiv preprint arXiv:1406.5823.

Defrancq, B., Plevoets, K. and Magnifico, C. (2015). "Connective Items in Interpreting and Translation: Where Do They Come From?" Yearbook of Corpus Linguistics and Pragmatics 2015. Bern: Springer, 195–222.

Ferraresi, A., Bernardini, S., Milicevic Petrovic, M. & Lefer, M-A. (2019). "Simplified or Not Simplified? The Different Guises of Mediated English at the European Parliament." Meta: Journal Des Traducteurs / Translators' Journal, 63(3), 717–738.

Forsyth, R. (2015). Formulib: Formulaic Language Software Library. Available at: http://www.richardsandesforsyth.net/zips/formulib.zip.

Forsyth, R. & Grabowski, Ł. (2015). "Is there a formula for formulaic language?" Poznań Studies in Contemporary Linguistics, 54 (1), 511-549.

Forster, P. (2001). "Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers". In: M. Bygate, P. Skehan, and M. Swain (Eds), Researching pedagogic tasks: Second language learning, teaching and testing. London: Longman, 75–93.

Hu, X., Xiao, R., & Hardie, A. (2016). "How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis". Corpus Linguistics and Linguistic Theory, aop [DOI: 10.1515/cllt-2014-0047]

Kajzer-Wietrzny, M. & Grabowski, Ł. (2021) "Formulaicity in constrained communication: An intermodal approach" In: Calzada, María & Sara Laviosa (eds.) 2021. Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection. MonTI 13, pp. 148-183. [DOI:10.6035/MonTI.2021.13.05][URL: https://www.e-revistes.uji.es/index.php/monti/article/view/6001 date of access: 1st July 2021]

Kruger, H., & Van Rooy, B. (2016a). "Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English". English World-Wide, 37(1), 26-57. https://doi.org/10.1075/eww.37.1.02kru

Kruger, H. & Van Rooy, B. (2016b). "Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans". Language Sciences, 56, 118-131.

Lanstyák, I. & Heltai, P. (2012). "Universals in Language Contact and Translation." Across Languages and Cultures 13 (1), 99–121. https://doi.org/10.1556/Acr.13.2012.1.6.

Nesi, H. (2012). "ESP and Corpus Studies". In: B. Paltridge & S. Starfield (Eds.), The Handbook of English for Specific Purposes. London: Wiley, 407-426.

Pęzik, P. (2018). Facets of prefabrication. Perspectives on modelling and detecting phraseological units. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Team, R.C., 2013. R: A language and environment for statistical computing.

Shlesinger, M. & Ordan, N. (2012). "More Spoken or More Translated? Exploring a Known Unknown of Simultaneous Interpreting." Target ,24(1), 43–60.

Schmitt, N. & Carter, R. (2004). "Formulaic sequences in action: An introduction". In: N. Schmitt (Ed.), Formulaic Sequences: Acquisition, Processing and Use. Amsterdam: John Benjamins, 1–22.

Siyanova-Chanturia, A. & Omidian, T. (2019). "Key issues in researching multi-word items". In S. Webb (Ed.), The Handbook of Vocabulary Studies. London, New York: Routledge,

Winter, B. (2019). Statistics for Linguists: An Introduction Using R. London: Routledge.

Wood, D. (2015). Fundamentals of Formulaic Language. London: Bloomsbury.

Building an SI corpus combining product and pre-process data of learners and professionals

Eva Klüber, Kerstin Kunz, Christoph Stoll
Ruprecht-Karls-Universität Heidelberg
eva.klueber@iued.uni-heidelberg.de, kerstin.kunz@iued.uni-heidelberg.de, christoph.stoll@iued.uni-heidelberg.de

Introduction

The aim of this paper is twofold: to present the architecture and on-going collation of a series of simultaneous interpreting (SI) subcorpora, integrated in the Heidelberg Conference Interpreting Corpus (HeiCIC), as well as the research in progress to be done on the core corpus. HeiCIC contains authentic speeches from LSP domains, simultaneous interpretations by learners and professionals in eight languages. The English-German core corpus is aligned with pre-process data. Our current research has two objectives: a) analysing semantic transfer from source to target text and b) correlating semantic transfer with pre-process data to determine which features reflect high-performance SI strategies.

Motivation

There are several aspects that set the corpus apart from other SI corpora: To date, no large, comparative learner/professional LSP corpus exists for SI, least for the language combinations in focus here. There are some learner corpora for Chinese – English (Leung and Yip 2013, Wei 2017), which are rather limited in size. Professional interpreter corpora such as EPIC, EPTIC and EPICG (Bernardini et al. 2018) focus on EU interpreting and are rather heterogeneous in terms of topic, register and level of technicality. Others, e.g. NAIST (Japanese – English) (Neubig et al. 2018), reflect interpreting environments for a lay audience, or incorporate other forms of interpreting, e.g. SIREN, which includes simultaneous interpreting with text and television interpreting (English – Russian) (Dayter 2018).

HeiCIC is designed to map authentic professional settings, where the highly technical nature of LSP and scientific conferences requires a structured, partially automated workflow for terminology and knowledge acquisition. Simultaneous interpreters integrate visual support material (CCT maps, see below) into their conference preparation in order to avoid cognitive overload while interpreting. Our corpus design is unique in that it aligns this pre-process data with both original speeches and interpreting output. This permits insights into advanced interpreting strategies used in LSP settings and thus process-related phenomena, while other corpora typically focus on product data (Gile 2002, Díaz Galaz 2015).

Corpus compilation

HeiCIC is collated mainly at the Heidelberg Conferences: scientists and experts present their research in a variety of LSP domains and send preparation material, which is used by students, young and seasoned professionals to prepare and then interpret from and into German, English, French, Italian, Spanish, Portuguese, Russian and Japanese. In total, the corpus comprises 83 hours of recorded original speeches and interpretations. Subcorpora differ in terms of formats available, languages included, LSP domains covered and level of interpreter expertise. The core corpus is a homogenous subpart containing several originals and multiple parallel interpretations per original by students, professionals with different levels of interpreter expertise, and transcripts (English – German) in selected LSP domains such as engineering, investor relations, AGMs. It currently contains more than 400,000 tokens and is constantly expanded as new recordings, transcripts and annotation layers are added.

We seek to follow basic principles of corpus compilation (Bernardini et al. 2018, Hansen-Schirra et al. 2012): Metadata are stored in a separate file for each transcript. They are structured into information

about speaker (e.g. gender, role, native language and language variety), interpreter (e.g. gender, experience, native language and language combination) and text (e.g. setting, language, register, topic and mode, text length in seconds and tokens) and allow for filtering according to these criteria. The transcription process includes several steps. Transcripts are generated using ASR combined with manual revision and aligned with the audio using WebMAUS (Kisler, Reichel and Schiel 2017). They are further processed with EXMARaLDA, which allows for alignment of several interpretations with one original speech (Schmidt and Wörner 2014). Our transcription guidelines are a slightly modified version of those for the GECCo Corpus (Kunz et al. 2011, Lapshinova et al. 2012). They include tags accounting for spoken language features, related to cognitive load in general (e.g. pauses, fillers, repairs) and to SI in particular (interpreter turns, incomplete sentences and grammatical errors), cf. Plevoets and Defranc 2016. The core part of the corpus contains automatic basic level annotations, such as tokenization, lemmatization and POS tagging. Transcripts, recordings and annotation layers are aligned with strategic pre-process data of interpreters (see below).

Research objectives and corpus analysis

One research objective pursued on a subpart of the corpus is the analysis of semantic transfer from source to target text, i.e. the reproduction of a message uttered in one language into another (Schjoldager 1995). Semantic units are identified in both texts based on semantic and grammatical information (Christoffels and de Groot 2005). A mapping of these units in source and target texts is attempted to categorise semantic transfer on a scale from omission, implicitation to explicitation and addition by assessing features in terms of their information content (Becher 2011, Hansen-Schirra et al. 2012) as well as shifts in position of the semantic units within sentence and text structure. Previous studies on SI have focused either on individual transfer phenomena such as explicitation or linguistic features such as cohesion markers (Gumul 2017, Kajzer-Wietrzny 2012). To our knowledge, a comprehensive analysis of semantic units and transfer categories in combination with the analysis of information structure has not been attempted so far.

In a second step, the properties of the interpretation output are correlated with pre-process data – preparations by interpreters, the CCT maps (content, concept and term maps). CCT maps contain chronological renderings of expected macrotopics reflecting textual function and skopos, ontological representations, and terminological organisations of topics. Furthermore, CCT maps integrate SI strategy cues relating predictions of source language problem triggers such as cognitive load conflicts and overruns (Seeber 2011-17) to efficient target language solutions (Stoll 2019). CCT maps are documented in several dimensions: In keeping with professional practice, conceptual and terminological information is combined into a single structure with different views for pre- and in-process phases (Stoll 2009, Fantinuoli 2012): While the pre-process view shapes terminology and expert knowledge into an ontological hierarchy (Rütten 2007, Will 2009), the in-process view lists terms, semantic fields and strategy cues in chronological order (Stoll 2009).

Our approach aims to determine which features in CCT maps can be identified as solution cues and therefore indicators of deliberate high-performance SI strategies as they correlate with the interpreter's output, thus proving process in product features. More specifically, correlating CCT maps, semantic transfer categories and interpreting output should yield information as to how predictions of source language problem triggers are marked and strategically related to efficient solution cues in CCT maps and how this is manifested in solutions in the interpretation output.

For instance, some SI strategy cues enhance speech production and monitoring relief. They integrate structural and semantic compression and nominalisation into the CCT Map, such as*: x has submitted the annual financial statement* being rendered as *Berichtsvorlage*. In this way, correlating semantic transfer in the product with pre-process data allows to identify particular subtypes of implicitation as conscious

interpreting strategies. Other strategies directly link source and target language semantic relations and are related to anticipation and memory relief.

Conclusion and outlook

Our research inverts the traditional errors-and-omissions-based approaches in empirical interpreting studies to establish an evidence-based, hierarchical typology of verifiable strategies of semantic, conceptual, lexical and strategic priming. Our findings may additionally serve to improve computer-aided interpreting. We plan to make parts of our corpus freely accessible for corpus-querying via a web interface such as CQPWeb for independent validation, validity and reliability of our research. Moreover, the corpus is well documented to permit research beyond our current focus in the future.

References

Becher, V. (2011). *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Hamburg: Universität Hamburg.

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. Bendalozzi & B. Defrancq (eds.) *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, 21-42.

Christoffels, I. K. & de Groot, A. M. B. (2005). Simultaneous interpreting: A cognitive perspective. In J. F. Kroll & A. M. B. de Groot (eds.) *Handbook of Bilingualism: Psycholinguistic Approaches*. New York: Oxford University Press, 454–479.

Dayter, D. (2018). Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM* 16(2), 241-264.

Díaz-Galaz, S. (2015). *La influencia del conocimiento previo en la interpretación simultánea de discursos especializados: Un estudio empírico*. PhD thesis, Universidad de Granada.

Fantinuoli, C. (2012). *InterpretBank - Design and Implementation of a Terminology and Knowledge Management Software for Conference Interpreters*. Berlin: epubli GmbH.

Gile, D. (2002). The Interpreter's Preparation for Technical Conferences: Methodological Questions in Investigating the Topic. *Conference Interpretation and Translation* 4(2), 7-27.

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

Hansen-Schirra, S., Neumann, S. & Steiner, E. (2012). *Cross-linguistic corpora for the Study of Translation: Insights from the Language-Pair English German*. Berlin: de Gruyter.

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

Kellett Bidoli, J. C. (2016). Methodological challenges in Consecutive Interpreting Research: Corpus analysis of notes. In C. Bendazzoli & C. Monacelli (eds.) *Addressing methodological challenges in Interpreting Studies Research*. Newcastle upon Tyne: Cambridge Scholars, 141-169.

Kisler, T., Reichel, U. D. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.

Lapshinova-Koltunski, E., Kunz, K. & Amoia, M. (2012). Compiling a Multilingual Spoken Corpus. In *Spoken Corpora and Annotation; Speech Technology and Data Bases. Proceedings of the VIIth GSCP International Conference*. Firence: Firence University Press.

Leung, S. M. E., & Yip, L. (2013). A bilingual corpus of interpreting students' performance. http://arts.hkbu.edu.hk/~engester/main.html.

Neubig, G., Shimizu, H., Sakti, S., Satoshi, N. & Toda, T. (2018). The NAIST Simultaneous Translation Corpus. In M. Russo, C. Bendalozzi & B. Defrancq (eds.) *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, 205-215.

Plevoets, K. & Defranc, B. (2016). The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies* 11 (2), 202-224.

Rütten, A. (2007). *Information and Knowledge Management in Conference Interpreting* (in German), Frankfurt: Lang.

Schjoldager, A. (1995). An Exploratory Study of Translational Norms in Simultaneous Interpreting: Methodological Reflections. *Hermes, Journal of Linguistics* 8(14), 65-88.

Schmidt, T. & Wörner, K. (2014). EXMARaLDA. In *Handbook on Corpus Phonology*. Oxford: Oxford University Press, 402-419.

Seeber, K. (2011). Cognitive load in simultaneous interpreting. Existing theories – new models. *Interpreting* 13(2), 176-204.

Seeber, K. (2013). Cognitive load in simultaneous interpreting: Measures and methods. *Target* 25(1), 18-32.

Seeber, K. (2017). Multimodal processing in simultaneous interpreting. In J. W. Schwieter & A. Ferreira (eds.) *The Handbook of translation and cognition.* New Jersey: Wiley Blackwell.

Stoll, C. (2002). Dolmetschen und neue Technologien. In J. Best & S. Kalina (eds.) *Übersetzen und Dolmetschen. Eine Orientierungshilfe*. Tübingen: Francke, 307-312.

Stoll, C. (2009). *Jenseits simultanfähiger Terminologiesysteme. Methoden der Vorverlagerung von Kognition im Arbeitsverlauf professioneller Konferenzdolmetscher*. Trier: WVT.
Stoll, C. (2019). Terminology Systems and Workflow Automation for Simultaneous Interpreters: CAI tools and Research within the HeiCIC Corpus (in German). *edition* 2019(1), 25-33.
Will, M. (2009). *Interpreting-Oriented Terminology Work* (in German). Tübingen: Narr.
Zhang, W. (2017). Chinese interpreting learner corpus construction and research: Theory and practice (in Chinese). *Chinese Translators Journal* 38(1), 53-60.

Relativizers as markers of grammatical complexity : A diachronic, cross-register study of English and German

Marie-Pauline Krielke
Saarland University
mariepauline.krielke@uni-saarland.de

Introduction

In the present paper, we aim to investigate grammatical complexity as a register feature of scientific English and German. Specifically, we carry out a diachronic comparison between general and scientific discourse in the two languages throughout 300 years (1600 - 1900) using relativizers as proxies to investigate the development of grammatical complexity. We ground our study in register theory (Halliday and Hasan, 1985), assuming that language use reflects contextual factors (i.e., field, tenor and mode), which contribute to the formation of registers (Quirk et al., 1985; Biber et al., 1999; Teich et al., 2016). The period between 1600 and 1900 is especially interesting, since academic disciplines and with them scientific discourse emerges (Görlach, 2008). Register theory assumes that different text classes not only differ from general language in topic or field, but also in terms of lexico-grammatical features reflecting tenor and mode. This has been shown in numerous corpus-linguistic studies (Biber, 1988, 1993, 2006, 2012). Teich et al. (2016) follow the hypothesis that the development of scientific language undergoes two parallel processes, specialization, and diversification. They show that over time scientific communication becomes increasingly expert-oriented, while the different scientific disciplines develop their own distinct set-up of lexico-grammatical features, distinguishing them from other disciplines. Specifically, for scientific English, previous research has shown a development towards higher lexical density (López-Couso et al., 2012; Biber, 2006; Biber and Gray, 2016; Degaetano-Ortlieb et al., 2016), while syntax becomes less complex indicating specialization of the scientific register (Halliday, 1988; Teich et al., 2016). In contrast to the findings for English, German diachronic studies have shown a strong Latin influence until the 18th century, resulting in a complex hypotactic syntax, with a considerable number and depth of embeddings made possible through more refined conjunctions. In later periods, however, a trend towards detangling this complex syntax can be observed (Admoni, 1990; Beneš, 1981; Habermann, 2011; Möslein, 1974). Based on the findings for the two languages, we assume that grammatical complexity may be a register feature shaping scientific discourse over time.

Hypotheses

In the present paper, we use relativizers as proxies to investigate the development of grammatical complexity for English and German. We look at three important aspects of grammatical complexity - syntactic intricacy, paradigmatic richness, and contextual predictability - pursuing the following hypotheses:

In scientific discourse

1. syntactic intricacy, as indicated by the use of relativizers and the number of relative clause embeddings within a sentence decreases in English and increases in German.
2. paradigmatic richness, indicated by the number of different relativizers decreases in English and increases in German.
3. contextual predictability of relativizers, i.e., relativizers appear in increasingly similar contexts.

Since our analyses affect two different languages situated in different socio- cultural contexts, we expect to find differences in the course of development of the scientific registers, notably, because scientific discourse became institutionalized in the UK primarily through the Royal Society of London in 1665, while for German scientific publications no such institution existed.

Data and Methods

For scientific writing, we use the Royal Society Corpus (RSC v4.0; Kermes et al. (2016)), consisting of the Proceedings and Transactions of the Royal Society of London covering the time from 1665-1869 with approx. 32 million tokens, including metadata (e.g., author, publication year) and linguistic annotation (e.g., tokens, lemmas, parts of speech, surprisal). For general English, we use the Corpus of Late Modern English Texts (CLMET v3.1; Diller et al. (2011)), spanning 1710-1920 with approx. 40 million tokens from several genres (e.g., narrative, drama), processed in the same way (TreeTagger, VARD). For German, data are taken from the scientific and general language subcorpora of Deutsches Textarchiv (DTA, Geyken et al. (2018)) respectively. Scientific German is represented with approx. 80 million tokens, general German with approx. 60 million tokens. To trace the development of grammatical complexity of the scientific register in the two languages, we focus on its respective features shown in Table 1.

Table 1: Features of grammatical complexity

Discourse Property	**Feature Category**	**Feature subcategory**	**Measure**
Grammatical complexity	Grammatical Intricacy	Freq. relativizers relativizers per sentence	Relative frequencies
	Paradigmatic richness	Relativizer paradigm	Entropy
	Contextual predictability	Probability of relativizers given their context	Surprisal

We make use of conventional frequency-based methods to account for syntactic intricacy indicated by the frequency of relativizers in the two registers as well as the number of relative clause embeddings within a sentence. To assess paradigmatic changes (growth or reduction), we use entropy "increase[ing] with a higher number of members of the paradigm as well as with greater similarity of the probabilities of the members" (Milin et al., 2009). To account for predictability of items in context, we use surprisal (Milin et al., 2009; Linzen and Jaeger, 2016). We analyze surprisal values of the different relativizers and inspect preceding 3gram part-of-speech sequences representing highly predictable contexts.

Preliminary Findings

Syntactic intricacy. Analyses of relativizer frequencies in English show a decrease in both registers with overall higher fpm in scientific language. For German, the general tendency is increase followed by decrease in both registers, however the peak in scientific language (1780) is higher and later than in general German (1750). The trends differ between the two languages but are similar between registers. Average number of relative clauses embedded per sentence coincides in both English registers with the trend in relativizer frequency. In both scientific and general German, relative clause embeddings per sentence abound between 1700 and 1750. Interestingly, in scientific German relative frequencies of relativizers peak later, suggesting a trend towards distributing relativizers more evenly across sentences.

Paradigmatic richness. For English, we find a striking reduction in entropy in scientific language (1.5 to 0.5), while entropy in general English stays stable (1). In German, we see an inverse development: In general German we find decreasing values of entropy (2 to 1.5), while in the scientific subcorpus entropy is fairly stable (2) and only decreases in the second half of the 19th c (1.6).

Contextual predictability. For scientific German, we see a general downward trend in surprisal for all different relativizers, indicating that relative clauses increasingly occur in similar contexts. For scientific English, surprisal of the main relativizer, *which*, stays fairly stable, while it increases for the pronominal adverbs on the decline. Qualitative analyses of the syntagmatic environments of relativizers show that in scientific English, *which* tends to increasingly occur in prepositional contexts (*the manner in which, some*

of which) expressing manner and quantification, while for German the most decisive factor for decreasing surprisal is prescriptive use of a comma preceding a relative clause.

References

Admoni, W. (1990). *Historische Syntax des Deutschen*. Niemeyer.

Beneš, E. (1981). Die formale Struktur der wissenschaftlichen Fachsprachen aus syntaktischer Hinsicht. In T. Bungarten (ed.) *Wissenschaftssprache*. München: Fink, 185–212.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. (1993). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities*, 26(1), 331–345.

Biber, D. (2006). University Language: A Corpus-based Study of Spoken and Written Registers, volume 23 of Studies in Corpus Linguistics. Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, D. (2012). Register as a Predictor of Linguistic Variation. Corpus Linguistics and Linguistic Theory, 8(1), 9–37.

Biber, D. and Gray, B. (2016). Grammatical Complexity in Academic English: Linguistic Change in Writing. *Studies in English Language*. Cambridge: Cambridge University Press.

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

Degaetano-Ortlieb, S., Kermes, H., Khamis, A., & Teich, E. (2016). An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In Suhr, C., Nevalainen, T. & Taavitsainen, I. (eds.), Selected Papers from Varieng – From Data to Evidence (d2e), *Language and Computers*. Brill.

Diller, H.J., De Smet, H., Tyrkkö, J. & Flach, S. (2011). A European database of descriptors of English electronic texts. *The European English Messenger* 19, 21–35.

Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C. & Wiegand, F. (2018). Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In *Germanistische Sprachwissenschaft um 2020*, volume 6. De Gruiter, 219–248

Görlach, M. (2008). *Text Types and the History of English*. Boston: De Gruiter Mouton.

Habermann, M. (2011). *Deutsche Fachtexte der Neuzeit. Naturkundlich-medizinische Wissensvermittlung im Spannungsfeld von Latein und Volkssprache.* Berlin/ Boston: De Gruyter.

Halliday, M.A.K. (1988). On the Language of Physical Science. In Ghadessy, M. (ed.) *Registers of Written English: Situational Factors and Linguistic Features*. London: Pinter, 162–177.

Halliday, M.A.K. & Hasan, R. (1985). *Language, Context and Text: A Social Semiotic Perspective.* Language and Learning Series. Geelong, VIC, Australia: Deakin University Press.

Kermes, H., Knappen, J., Khamis, A., Degaetano-Ortlieb, S. & Teich, E. (2016). The Royal Society Corpus: Towards a high-quality corpus for studying diachronic variation in scientific writing. In *Proceedings of Digital Humanities* (DH).

Linzen, T. & Jaeger, F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science* 40(6), 1382–1411.

López-Couso, M., Aarts, B. & Méndez-Naya, B. (2012). Late Modern English Syntax. In Bergs, A. & Brinton, L.J. (eds.) *Historical Linguistics of English*. Mouton de Gruyter, 869–887.

Milin, P., Kuperman, V., Kostic, A., & Baayen, H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in Grammar: Form and Acquisition*, 214–252.

Möslein, K. (1974). Einige Entwicklungstendenzen in der Syntax der wissenschaftlich- technischen Literatur seit dem Ende des 18. Jahrhunderts. *Zur Geschichte der deutschen Sprache und Literatur* 94, 156–198.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Teich, E., Degaetano-Ortlieb, S.; Fankhauser, P., Kermes, H. & Lapshinova-Koltunski, E. (2016). The Linguistic Construal of Disciplinarity: A Data-mining Approach Using Register Features. Journal of the Association for Information Science and Technology (JAIST) 67(7), 1668–1678.

Using corpora for post-editing neural MT in highly specialised domains: the case of complex noun phrases

Natalie Kübler[1], Hanna Martikainen[2], Alexandra Mestivier[1] and Mojca Pecman[1]
Université de Paris[1], École Supérieure d'Interprètes et de Traducteurs (ESIT), Université de la Sorbonne Nouvelle Paris 3[2]
nkubler@eila.univ-paris-diderot.fr, hanna-julia.martikainen@sorbonne-nouvelle.fr, alexa.volanschi@gmail.com, mpecman@eila.univ-paris-diderot.fr

The present study focuses on the use of specialised comparable corpora for post-editing machine translation in LSPs, a recent development to the methodological framework which was introduced in our translation syllabus 20 years ago and which has been greatly improved over the years (Kübler 2003, Kübler 2011, Kübler et al. 2016, Kübler et al. 2018). This framework for teaching specialised translation to master's students involves a wide range of competences. Students have classes in corpus linguistics, corpus query tools, terminology and specialised translation. An experimental protocol with the Earth and Planetary Sciences (EPS) department was created to put the framework into practice. Our translation trainees first collaborate with EPS students on the terminology of highly specialised English-language articles and then translate excerpts from the articles into French. Since 2013, our team has been assessing the impact of corpus use on students' translations through the study of learner corpora. Every year, we assemble two translator learner corpora containing translations produced 1) without using corpora, and 2) using all available resources, and specifically the specialised comparable corpora the students compile during terminology analysis. Both student translations sub-corpora are annotated with the MeLLANGE error typology (Castagnoli et al. 2011) on a Brat server. A previous comparison of the two shows a positive impact of corpus use on some error categories, but limited impact on other errors categories, among which Distortion (Kübler et al. 2016, 2018).

For the past two years, a further component has been introduced in our framework, i.e. postediting MT output using corpora. As stated in the European Master's in Translation (EMT) Competence Framework (2017), "the ability to interact with machine translation in the translation process is now an integral part of professional translation competence" - hence the need to train translation students for this task. The development of neural MT, which is state-of-the art in machine translation today, clearly increases fluency compared with the previous generation of statistical engines (Toral & Sanchez-Cartagena 2017), but shows mixed results when evaluated by humans (Castilho et al. 2017, Toral & Sanchez-Cartagena 2017). Even though professional translators still seem to mistrust neural MT, Scansani et al. (2019) show that students in translation show no difference in trust when revising human translation or post-editing MT. Most of our trainee translators are familiar with MT, and have a rather positive image of the technology (Kübler et al. in press). Previous comparison of human-translated & post-edited texts in two highly specialised domains shows that, while the intervention of neural MT globally results, for our students, in improvements in both productivity and quality, it also leads to specific errors (Martikainen & Mestivier 2019, Kübler et al. in press).

Within this framework, we have previously studied the use of corpora in the translation of complex noun phrases (CNP), which are characteristic of the highly specialised texts our students work on. Our previous results show that CNPs are a frequent source of distortion in translation. In the present research, we focus on the impact of neural MT on the translation of CNPs in highly specialised domains and on the use of corpora in the post-editing process of this type of construction in particular.

During the 2018-2020 academic years, we have revised our procedure for evaluating the impact of corpus use on translation: error-annotating student translations now includes annotating MT output and post-

edited texts. Students thus submit several texts: a) the translation of a 250 word fragment for which they may only use available term-bases and Internet searches b) a second version of this translation improved by using corpora, c) the translation of a 250 word fragment achieved by post-editing MT output (aligned with the MT output) and d) and improved version achieved by using corpora. The comparison of the four text-types and students' comments on the translation process allow us to formulate tentative answers for several research questions: How well do MT systems perform on the translation of CNPs? Are translation error typologies adequate for annotating MT output and post-edited texts? What are the differences between MT and human translation where CNPs are concerned and how does the former influence the latter when post-editing? Is corpus use efficient in post-editing MT when translating CNPs (studied mainly through student comments)? Can we devise better adapted teaching material to draw student's awareness to the CNP translation difficulties?

Regarding neural MT performance on complex NPs, preliminary results suggest that the calque solutions produced by the generic MT engine, although adapted in some contexts, frequently require some amount of post-editing in this highly specialised domain. In some cases, minimal post-editing effort is required, as for the segment "an exhumed serpentinite-sediment contact", for which MT output is "un contact serpentine-sédiment exhumé". A single edit operation would be necessary here, to replace the erroneous term "serpentine" by the correct one, "serpentinite". Trainee translators, however, sometimes over-edit such calque translations of CNPs, specifically when they contain specialised terms derived from general vocabulary or from other specialised domains, such as "contact" and "exhumé" in this example, both of which can be observed in the specialised comparable corpora. In other instances, MT calque solutions for complex NPs require extensive post-editing. This is the case for the segment "hydrothermally altered rocks", for which the translation produced by the generic MT engine ("roches altérées *hydrothermiquement") contains a non-existent word, or a 'hallucination' (Moorkens 2020), namely "hydrothermiquement". When the complex NP is composed of more familiar elements, as is the case here, students tend to under-edit MT output. In this second example, replacing the erroneous term by the correct one ("hydrothermalement") is not sufficient for producing an acceptable translation. Research in the specialised comparable corpora shows that, in this instance, it is necessary to unpack the compact construction by adding either a verbal construction (e.g. "roches ayant subi une altération hydrothermale") or a prepositional phrase (e.g. "roches altérées par hydrothermalisme").

These preliminary results thus consolidate our previous results on the usefulness of comparable corpora for LSP translation, and confirm the need to train translation students on using corpora for post-editing as well.

References

CASTILHO, Sheila, MOORKENS, Joss, GASPARI, Federico, *et al*. Is neural machine translation the new state of the art?. *The Prague Bulletin of Mathematical Linguistics*, 2017, vol. 108, no 1, p. 109-120.

EMT Competence Framework, 2017

KÜBLER, Natalie 2003. 'Corpora and LSP Translation.' in F. Zanettin, S. Bernardini and D. Stewart (eds), *Corpora in Translator Education*, 25–42, Manchester: St Jerome Publishing. Republished 2014 Routledge.

KÜBLER, Natalie 2011. Working with different corpora in translation teaching. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston, *New Trends in Corpora and Language Learning*. London: Continuum, 2011, p. 62-80.

KÜBLER, Natalie, MESTIVIER, Alexandra, PECMAN, Mojca & ZIMINA, Maria 2016 'Exploitation Quantitative de corpus de traductions annotés selon la typologie d'erreurs pour améliorer les méthodes d'enseignement de la traduction spécialisée', Actes des 13es Journées internationales d'Analyse statistique des Données Textuelles, 731–41, 7–10 June, Nice, France.

KÜBLER, Natalie, MESTIVIER Alexandra & PECMAN, Mojca 2018. 'Teaching Specialised Translation through Corpus Linguistics: Quality Assessment and Methodology Evaluation by Experimental Approach', *Meta: Journal des Traducteurs / Meta: Translators' Journal*, 63 (3): 806–24.

KÜBLER, Natalie, MESTIVIER, Alexandra et PECMAN, Mojca, in press. Using Comparable Corpora for Translating and Post-Editing Complex Noun Phrases in Specialised Texts: Insights from English-to-French Specialised Translation. In : GRANGER, Sylviane et LEFER, Marie-Aude (éd.), *Extending the Scope of Corpus-Based Translation Studies*. S.l. : Bloomsbury Academic. Bloomsbury Advances in Translation. ISBN 978-1-350-14325-8.

MARTIKAINEN Hanna. & MESTIVIER Alexandra. *Les outils de traduction nouvelle génération: quel effet sur la qualité des textes traduits ?* Traduction & Qualité: biotraduction et traduction automatique, Université de Lille, 31 janvier 2020.

MOORKENS Joss. Neural Machine Translation and Polarisation of the Translation Workplace. #TQ2020 Traduction & Qualité : biotraduction et traduction automatique, Université de Lille, France, 31/01/2020.

TORAL, Antonio et SÁNCHEZ-CARTAGENA, Víctor M. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*, 2017.

SCANSANI, Randy, BERNARDINI, Silvia, FERRARESI, Adriano, *et al.* Do translator trainees trust machine translation? An experiment on post-editing and revision. In : *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. 2019. p. 73-79.

Corpus-based study applications: bridging the gap between research and the marketplace

Belén López Arroyo, Leticia Moreno Pérez
Universidad de Valladolid
mariabelen.lopez@uva.es, leticia.moreno@uva.es

There has been a shift in the center of gravity in linguistic concerns over the last thirty years (Swales, 1990, 2004; Lee, 2001; Lewin et al., 2001 among others), from a focus on language as a set of syntactic structures in isolation to a focus on language as a set of functional resources in use (Scott & Thompson, 2000: 1). In this sense, the combination of Corpus-Based Studies and Contrastive Functional Analysis has provided a sound basis for cross linguistic description (Johansson, 2007; Rabadán Álvarez, 2007, 2008; Biber & Conrad, 2009). To this day, the purpose of most studies is to provide input to applied disciplines such as Foreign Language Teaching or Translation Studies. Thus, the alliance between Contrastive Functional Analysis and Corpus-Based Studies is supposed to provide an interface between theory and application. However, these cross-linguistic studies do not seem to have "successfully bridged the gap between theory and practice" (Rabadán Álvarez, 2008: 104), since most of them just report "raw descriptive data that may or may not have strong implications for applied (…) activities" (Rabadán Álvarez, 2008: 105).

In this era of globalization, not only specific discourse communities but also societies in general are demanding the development of useful and usable tools to improve crosslinguistic communication (Quesenbery, 2001; Kreitzberg and Little, 2009; Turner et al., 2019). In other words, we need to move from the description of linguistic resources to the prescription of useful and usable guidelines with "descriptively correct possibilities" (Rabadán Álvarez, 2008: 114) that satisfy the professional's or the scientist's language needs. The ACTRES research group (Spanish acronym for Contrastive Analysis and Specialized Translation), aware of this situation, has as one of its primary aims the development of tools that will allow a non-native speaker to write texts in a foreign language; these tools are envisaged as language-bound (in this case English-Spanish), computer-friendly applications which are restricted to specific scientific or professional genres of interest.

This study presents the steps taken and the semi-automatic tools developed to produce these computer applications called 'generators', which are based on the results of Descriptive and Corpus-Based Studies carried out within the framework of the research group. The purpose of these generators is to offer the user a reliable and ready-to-use tool to write, in a different language, a professional genre that the target discourse community can recognize and identify with. Our research has resulted in a series of patented tools which are already available on the market. The wine tasting notes generator will be the one used in this study to illustrate the necessary steps for the creation of these tools, as well as its usefulness.

References

Biber, D. and S. Conrad. 2019. *Register, Genre & Style*. Cambridge: Cambridge University Press.

Johansson S. 2007. *Seeing Through Multilingual Corpora*. Antwerp: John Benjamins.

Kreitzberg, Ch. and A. Little. 2009. "Usability in practice: Useful, usable and desirable. Usability as a core development competence". *MSDN magazine*. Available at http://msdn.microsoft.com/en-us/magazine/dd727512.aspx (accessed 21 November 2019).

Lee, D. Y.W. 2001. "Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle". *Language Learning & Technology*, 5 (3): 37-72.

Lewin, B. and L. Young. 2001. *Expository Discourse: A Genre Based Approach to Social Science Research Texts*. London: Continuum.

Moreno Pérez, L. and B. López Arroyo. 2021. "Atypical Corpus-Based Tools to the Rescue: How a Writing Generator Can Help Translators Adapt to the Demands of the Market", *Monti 13*.
Quesenbery, W. 2001. What Does Usability Mean: Looking Beyond 'Ease of Use'. Proceedings of the 48th Annual Conference, Society for Technical Communication, 2001. Available at https://www.wqusability.com/articles/more-than-ease-of-use.html, (accessed 21 November 2019).
Rabadán Álvarez, R. 2007. "Divisions, description and applications- The interface between DTS, Corpus-Based research and contrastive analysis". In Y. M. Gambier, M. Schlesinger and R. Stolze (eds.) 2007 *Doubts and Directions in Translation Studies*, 237252. Antwerp: John Benjamins.
------ 2008. "Refining the idea of 'applied extensions'". In A. Pym, M. Schlesinger & D. Simeoni eds 2008 *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*, 103-118. Antwerp: John Benjamins.
Scott, M. and M. Thompson. 2000. "Introduction. Why 'patterns of text'?" In M. Scott and M. Thompson (eds) 2003 *Patterns of Text. In honour of Michael Hoey*, 1-7. Antwerp: John Benjamins.
Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
------ 2004. *Research Genres*. Cambridge: Cambridge University Press.
Turner, A. M. et al. 2019. "Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study." *JMIR public health and surveillance*, vol. 5,1 e11171. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6369422 (accessed 21 November 2019).

Scrutinizing gerunds. A multifactorial perspective on unique items.

Charlotte Maekelberghe, Isabelle Delaere
KU Leuven
charlotte.maekelberghe@kuleuven.be, isabelle.delaere@kuleuven.be

The past few years, the discipline of Corpus-based Translation Studies has undergone a number of significant developments (De Sutter, Lefer & Delaere 2017; De Sutter & Kruger 2018). The empirical turn and its methodological innovations, such as the implementation of multivariate statistics, incentivized a shift from purely frequency-based analyses of translation universals to more complex and probabilistic assessments of language features (De Sutter & Lefer 2019). From a theoretical perspective, the study of translations has been embedded in a broader model of "constrained communication", which posits that linguistic features typical of translated language can also be found in other types of mediated language, such as learner or non-native varieties (Lanstyák & Heltai 2012; Kruger & Van Rooy 2016).

The present paper wishes to apply these recent insights to the study of the English gerund, which is a particular type of *unique item*, i.e. a linguistic item that lacks a straightforward translation equivalent in the target language (see also Halverson 2003; Tirkkonen-Condit 2004). The English gerund is characterized by its formal hybridity, combining clausal internal syntax with the external distribution of a noun phrase.[8] Importantly, gerunds display functional hybridity as well, ranging from more clausal instances to uses that are more typically nominal. These distinct functional profiles have been shown to correlate with particular coreferentiality patterns and syntactic functions (Maekelberghe 2020), whereby gerunds which are controlled by a matrix clause participant and which occur in adverbial positions, as in (1a), display more functional overlap with clausal constructions, while "uncontrolled" gerunds–often in subject position–are more reminiscent of regular abstract nouns, as in (1b) (see also Langacker 1991: 25).

(1) a. We got faster by *reducing corporate staff*. (CroCo, English Original)
b. But *building a common culture* is a huge challenge that needs to be actively shaped and carefully managed. (DPC, English Original)

In this study, we propose a more contextualized and multifactorial approach to the study of unique items. Rather than comparing the frequency of gerunds across different varieties, we wish to investigate the following research questions: (i) How can we map out the distinct functional profiles of gerunds, in terms of preferred configurations of clausal functions, coreferentiality patterns and structural complexity and (ii) how do these profiles differ across various (constrained) language varieties?

Concretely, we will analyze data from CroCo (Hansen-Schirra et al. 2012) and DPC (Macken et al. 2011), which contain data for the language pairs English-German and English-Dutch respectively. In order to ensure maximum comparability, data from the following subcorpora were selected:

(Sub)corpus	Selected text types
CroCo: ● original English (Cr_EO) ● English translated from German (Cr_ET)	*Essays, Fiction, Instructive texts, Popular-scientific texts, Letters to shareholders, Speeches*

[8] We are hence not concerned with the fully nominal gerund variant, as in e.g. *the building of a common culture*.

DPC: ● original English (DPC_EO) ● English translated from Dutch (DPC_ET)	*Proceedings of parliamentary debates, Fiction, Instructive texts, Journalistic texts on science, Yearly reports and self-presentations, Official speeches*

After extracting all verbal *-ing* forms and manually filtering for instances of gerunds (hence excluding participial or fully nominal *-ing* forms), we compiled a data set comprising 11,106 gerunds in total. The dataset was then manually annotated for a number of language-internal and language-external factors, including genre (based on the metadata), clausal function, coreferentiality and structural complexity (in terms of number of words). In order to map out the functional usage profiles, we applied a Hierarchical Configural Frequency Analysis (henceforth HCFA; von Eye 1990, Gries 2004), an exploratory statistical technique which identifies configurations of values that have a tendency to co-occur (Hilpert 2009: 45).

The results reveal four significant configurations across varieties and genres: (i) controlled gerunds in prepositional slots, (ii) uncontrolled gerunds in subject position, (iii) uncontrolled 'independent' gerunds, which are not embedded in the clause structure, and (iv) gerunds with explicit subject in direct object position. Gerunds as in (iii) are especially prevalent in instructive texts, while those in (iv) are significantly associated with fictional texts.

i. (…) he started by *steering AG Continental into safe waters* (…). (Cr_ET, letters to shareholders)
ii. *Harnessing China to the global system* is crucial. (Cr_EO, essay)
iii. "*Creating a thin client response file (Windows)*" on page 67 (DPC_ET, instructive texts)
iv. You don't mind *me calling you Vivien*? (DPC_EO, fiction)

However, the distribution of these configurations differs in the two corpora and varieties. In CroCo, we find that, although gerunds occur equally frequently in original and translated English, configurations as in (i) are more frequent in translated English, while those in (ii) occur more frequently in original English. Interestingly, the opposite observation holds for DPC. Gerunds are significantly more frequent in DPC translated English, where configurations like (ii) and (iii) are more dominant. Those in (i), in contrast, are associated with original English. In DPC, gerunds in original English are also significantly more complex than those in translated English, while no significant difference was found in CroCo.

The contradictory evidence from CroCo and DPC raises a number of questions, which will be addressed in our paper. Firstly, to what extent is the observed variation in preferred configurations constrained vs. free? Secondly, can differences between the two corpora be attributed to the different language pairs, or rather to structural differences between the genres included in the (sub)corpora, or both? Therefore, a thorough analysis of the (meta)data will be carried out so as to paint a picture which is as detailed as a corpus-based approach will allow. Ultimately, the aim of this paper is to reveal probabilistic tendencies, rather than universal truths (see also Neumann 2021). We argue this approach might lead to hypotheses regarding the interaction between internal and external norms in translation and language production in general (Halverson & Kotze 2021).

References

De Sutter, G., M-A. Lefer & I. Delaere (eds.). (2017). *Empirical Translation Studies. New Methodologies and Theoretical Traditions.* Berlin: Mouton de Gruyter.

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.

Gries, S.Th. (2004). HCFA 3.2 — A Program for hierarchical configural frequency analysis for R for windows.

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

Halverson, S. L. & Kotze, H. (2021, February 25). *Social Aspects.* [Panel presentation] TRICKLET Workshop 2021. Model Building in Empirical Translation Studies.

Hansen-Schirra, S., Neumann, S. & Steiner, E. (2012). *Cross-Linguistic Corpora for the Study of Translations. Insights from the language pair English-German*. Berlin: Mouton de Gruyter.
Hilpert, M. (2009). The German mit-predicative construction. *Constructions and Frames*, 1(1): 29–55.
Kruger, H. & De Sutter, G. (2018). Alternations in contact and non-contact varieties : reconceptualising that-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior*, 1(2), 251–290.
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.
Langacker, R.W. (1991). *Foundations of Cognitive Grammar 2: Descriptive Application*. Stanford: Stanford University Press.
Lanstyák, I. & P. Heltai. (2012). Universals in language contact and translation. *Across Languages and Cultures* 13(1), 99–121.
Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus. *Meta*, 56(2), 374–390.
Maekelberghe, C. (2020). *The Present-day English Gerund System: A Cognitive-constructionist Account*. Berlin: Mouton de Gruyter.
Neumann, S. (2021, February 25). *Social Aspects.* [Panel presentation] TRICKLET Workshop 2021. Model Building in Empirical Translation Studies.
Tirkkonen-Condit, S. (2004). Unique items – Over- or under-represented in translated language? In P. Kujamäki, & A. Mauranen (Eds.), *Translation universals. Do they exist*? (pp. 177–184). Amsterdam: Benjamins.
von Eye, A. (1990). *Introduction to configural frequency analysis: the search for types and antitypes in cross-classifications*. Cambridge: Cambridge University Press.

Use of English loanwords containing V-ING type forms in German, Spanish, French and Italian: a corpus-based study of the European parliament debates.

François Maniez, María Belén Villar Díaz, Farge Sylvain
Centre de recherche en linguistique appliquée (CeRLA) - Université Lumière Lyon 2
francois.maniez@univ-lyon2.fr, maria-belen.villar-diaz@univ-lyon2.fr, sylvain.farge@univ-lyon2.fr

Words starting with a verb root and ending with the *-ing* morpheme feature prominently among words borrowed from English in many Indo-European languages, and the rising popularity of the *-ing* morpheme has been attributed by some scholars (Picone 1996) to its nominalizing syntactic function. Such borrowings are frequently followed by the creation of equivalents coined by using native words in the receiving language, and occasionally by their inclusion in standard dictionaries (cf. the case of *brainstorming* and *remue-méninges* for French, Humbley 2008). Such neologisms occasionally present as hybrid borrowings (*surbooking* for *overbooking* in French) or pseudo-Anglicisms such as *mailing*.

Using the multilingual alignments of the Europarl corpus (Tiedemann 2009), which includes all of the European Parliament debates between 1996 and 2003 in 11 European languages, we studied the use of terms and expressions containing a word beginning with a verb base and ending with the *-ing* morpheme (e.g. *benchmarking*) in German; Spanish, French and Italian.

In all four languages concerned we found a strong tendency to borrow terms consisting of a single word but a greater resistance to multi-word expressions, even if some of them (*level playing field*, *naming and shaming*, *paying but not playing*, *no trade without tracking, sparring partners*) are occasionally translated literally in the corpus, a fact which might be due to the difficulties such expressions posed for simultaneous interpreters.

All borrowings from English ending with the *–ing* morpheme were extracted from the corpus in all four languages and those which were used at least ten times in one of the four languages were analyzed quantitatively. From a quantitative point of view, German seemed to be the language with most such borrowings from English, followed by Italian, whereas French and Spanish showed relatively more resistance to *-ing* loanwords than those two languages, perhaps (in the case of French) due to the role of government institutions such as the *Délégation Générale à la Langue Française*, whose goal is to provide native equivalents to foreign loanwords. Italian seemed to exhibit a lesser degree of resistance to *-ing* forms than Spanish and French for many such forms (*roaming*, *doping*, *overbooking*, *trading*), while some of them were even used as translation equivalents for other English words (e.g. *mobbing* for *harassment*).

As the Europarl corpus reflects language use that is two decades old, we also occasionally compared the figures we obtained for the most frequently used *-ing* forms in our corpus with data from more recent corpora, such as the Google Books corpus and the large Web corpora compiled by the University of Leipzig over the last decade.

Our study also revealed that in languages where terms ending in *–ing* were not consistently borrowed in translation, a fair amount of variation could be observed in the translation equivalents that were chosen, possibly due to the fact that the terminology of the relevant domain was not yet stabilized in those languages at the turn of the century. For instance, the term *e-learning* was massively borrowed in German and Italian, but much less so in the other two languages under study, prompting a high degree of variation in translation equivalents both in French and Spanish.

The extent of the borrowing phenomenon also seemed to be occasionally domain-dependent, as many loanwords seemed to relate to the domain of economy and finance (*rating, dumping, trading, holding, frontloading, pricing*), as well as technology and communications (*caching*, *roaming, remailing, spamming, unbundling,*).

From a quantitative point of view, taking into account all of the *–ing* forms used in the four languages under study that were used at least twice in the corpus confirms the original findings obtained when factoring in only the forms that were used ten times or more. There are 4183 tokens for such forms in Spanish and while a slightly higher percentage is observed in French (19,4%), the difference is much greater for Italian (68,8%) and German (116,2%).

From a qualitative point of view, a number of problems complicate the comparison process between the four languages involved when dealing with the Europarl corpus. One is orthographic variation (e.g. occasional absence of hyphenation in such words as *front-loading*), as well as compounding, especially in German. The word *screening*, for instance, numbers 154 tokens in German, but it is also used in 14 *ing* ending compounds totaling 57 tokens (*Mammographiescreening, Brustkrebs-Screening, KrebsScreening, Brust-Screening,* etc.) and is part of 72 other compound tokens either including the word *Screening* (*Brustkrebs-Screening-Programme*) or beginning with it (*Screening-Verfahren*). Another problem frequently encountered concerns some cases in which the loanword does not include the *–ing* morpheme and thus escapes automatic detection (We have web **streaming** and I am sure the interpretation will be there. Wir haben ja **Webstream** , die Verdolmetschung wird sicherlich zu finden sein.).

Finally, many uses of *–ing* loanwords refer to technical terms, so that their first uses are often found in a context in which the word is defined by the speaker, which often prompts literal use of the English word in translation:

EN : […] particularly with regard to the article on copying for technical purposes, or **caching**, […]
DE : […] vor allem in Bezug auf die Artikel über die technische Vervielfältigung, das so genannte **Caching**, […]
ES : […] sobre todo acerca del artículo relativo a la copia técnica, o sea, el llamado **caching**, […]
FR : […] en particulier en ce qui concerne l'article relatif à la copie technique, que l'on appelle également **caching**, […]
IT : […] che riguardano innanzitutto l'articolo relativo alla copia tecnica, ovvero il cosiddetto **caching**, […]

Another issue that needs to be investigated is the influence of the language originally used by the speaker. It thus seems that when the source language is not English (as in the example above), there is a higher probability that the *–ing* word will be translated literally, possibly because it is already identified as a loanword by translators.

References

ALVAR EZQUERRA M. (1995), *La formación de palabras en espagnol*, Madrid, Arco Libros. BARREAU J.-L. (2001), Des emprunts de l'espagnol péninsulaire aux autres langues européennes, *Cahiers de l'Institut de Linguistique de Louvain*, Vol. 27 (3-4), 89-112.

BENARDI, R. L. L'italien des institutions publiques, une langue bien perméable aux anglicismes. *N 13–décembre 2014*, 16.

BISTARELLI, A. (2008). L'interferenza dell'inglese sull'italiano. *inTRAlinea*, *10*, 1-11.

CARTIER, E. et al. (2018) "Détection automatique, description linguistique et suivi des néologismes en corpus: point d'étape sur les tendances du français contemporain." *SHS Web of Conferences*. Vol. 46. EDP Sciences.

FURIASSI, C., PULCINI, V., et GONZÁLEZ, F. R. (ed.) (2012), *The anglicization of European lexis*. John Benjamins Publishing,.

GROSSMANN M., RAINER F. (dirs.) (2004), *La formazione delle parole in italiano*, Tübingen, Niemeyer.

HUMBLEY J. (2008), Emprunts, vrais et faux, dans le *Petit Robert 2007*, In PRUVOST J. (dir.), *Les journées des dictionnaires de Cergy : Dictionnaires et mots voyageurs*. *Les 40 ans du Petit Robert, de Paul Robert à Alain Rey*, 221-238, Herblay, Éditions des Silves.

KNOSPE, S. (2015). Pseudoanglicisms in the language of the contemporary German press, in Furiassi, Cristiano & Gottlieb, Henry (Hg.): *Pseudo-English. Studies on False Anglicisms in Europe.* Berlin, New York (de Gruyter), 99-122.

MAKRI-MOREL J. (2009), *La création lexicale en espagnol péninsulaire contemporain : étude néologique, typologie des procédés et réflexions*, Lille, ANRT.

MIRANDA J. A. (1994), *La formación de palabras en español*, Salamanca, Colegio de España.

PICONE M. D. (1996), *Anglicisms, Neologisms and Dynamic French*, Amsterdam, Benjamins.

TIEDEMANN J. (2009), News from OPUS – A collection of multilingual parallel corpora with tools and interfaces, in NICOLOV N., ANGELOVA G., MITKOV R. (eds.), *Recent Advances in Natural Language Processing V: Selected papers from RANLP 2007*, 237-248, Amsterdam, Benjamins.

TOURNIER J. (1991), *Précis de lexicologie anglaise*, Paris, Nathan.

VARGA, D., OREŠKOVIĆ DVORSKI, L., & BJELOBABA, S. (2011). English loanwords in French and Italian daily newspapers. *Studia Romanica et Anglica Zagrabiensia: Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l'Université de Zagreb*, *56*, 71-84.

Internet sources

Cent termes français du vocabulaire technique recommandés par la CGTN (2004, MCC, DGLFLF), www.dglflf.culture.gouv.fr/publications/vocabulaires/100termes.pdf

Leipzig Corpora Collection, https://corpora.uni-leipzig.de

OPUS, the open parallel corpus, http://opus.lingfil.uu.se/

Vocabulaire de l'économie et des finances (2006), http://fr.scribd.com/doc/120507304/Vocabulaire-del%E2%80%99economie-et-des-finances

The Gravitational Pull Hypothesis and imperfective/perfective aspect in Catalan translated and non-translated literary texts

Josep Marco[1], Llum Bracho Lapiedra[2], Gemma Peña Martínez[2]
Universitat Jaume I[1], Universitat Politècnica de València[2]
jmarco@uji.es, llumbra@idm.upv.es, gpenya@idm.upv.es

The aim of this paper is to test out the Gravitational Pull Hypothesis (GPH) on the imperfective/perfective aspect distinction. The study draws on the English- and French-Catalan sub-corpora (EN-CA and FR-CA, respectively) and the non-translation component (NTR) in COVALT – a multilingual corpus made up of the translations into Catalan of narrative works originally written in English, French, and German published in the autonomous region of Valencia from 1990 to 2000, together with their corresponding source texts. A comparable component of Catalan non-translations was later added.

The GPH was put forward by Halverson (2003, 2017) as an attempt to bring together various alleged properties of translated text, such as over- and under-representation of target language typical features. It posits three potential causes of translational effects: patterns of salience or prototypicality, which are target language internal (factor 1); conceptual structures/representation of the source language item, which are related to the structure of the source language (factor 2); and patterns of connectivity, which reflect relationships between the source and the target languages (factor 3). One effect is predicted for each potential cause, or factor. The effect of factor 1 will be over-representation; the effect of factor 2 will be over-representation too; and the effect of factor 3 may be over- or under-representation.

In the research reported on in this paper, the imperfective/perfective aspect distinction was chosen as a testing ground for the hypothesis because it is grammaticalised in Catalan and French but not in English. This is best seen in the past tenses. Catalan, like French, has two simple tenses with past reference: the preterite (which in Catalan can be synthetic or periphrastic) and the imperfect past. The preterite "combines past time with perfective aspect" (Wheeler et al. 1999: 343); the imperfect "is defined by two features: the past-ness and the extension over time of the event or state referred to" (Wheeler et al. 1999: 346). In the English verb system, the only simple tense with past reference is the simple past.

Particular hypotheses in the framework of the GPH need to take into account the relative salience of the elements constituting a semantic network. Drawing on claims made by several authors (Comrie 1976; Binnick 1991: 296; Maingueneau and Salvador 1995: 42; Pérez Saldanya 2002: 2579), it is assumed in this study that the preterite is more salient than the imperfect in narrative environments. As to the network involving these two simple past tenses in Catalan and their matching English forms, two facts receive special attention. Firstly, an English simple past can be construed as a perfective or an imperfective event and translated into Catalan as a preterite or an imperfect, respectively. And secondly, "English constructions like 'I used to go', 'I was going', and 'I would (habitually) go' will almost always correspond to the Catalan imperfect" (Wheeler et al. 1999: 346). A preliminary network can be established, then, on the basis of the forms just mentioned, as in Figure 1. On the other hand, the network involving the two simple past tenses in Catalan and their matching French forms must perforce look different, as French also makes the aspectual distinction in the past. In French, the *passé simple* is the tense prototypically used in *récit* (narrative). It has all but vanished from oral discourse, where the *passé composé* is used instead. It presents an event from a synthetic viewpoint as a limited whole with no links to the present. The *imparfait*, on the other hand, offers an internal view of the event with no consideration of temporal limits. Therefore, a much higher degree of overlap is to be expected between the French and the Catalan systems. The bilingual network might look like that depicted in Figure 2.

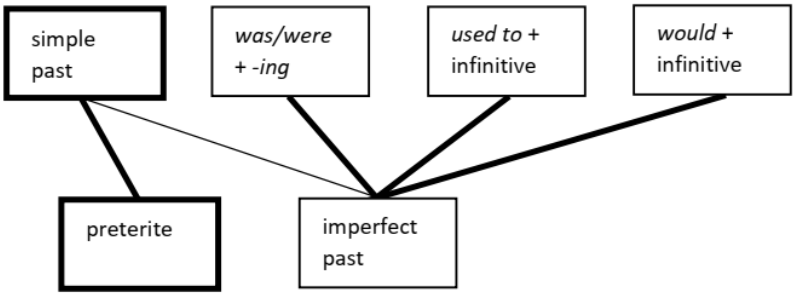


Figure 1. Network for the Catalan preterite and imperfect past and their English triggers

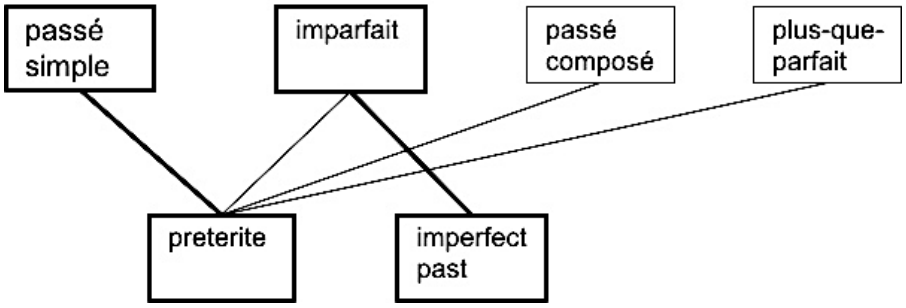


Figure 2. Network for the Catalan preterite and imperfect past and their French triggers

Such networks enable us to make the following predictions about the frequency of the two Catalan tenses:

1. The Catalan preterite will be over-represented in translations from English as compared to Catalan non-translations – and as a corollary the imperfect will be under-represented.
2. The frequency of the Catalan preterite in translations from French and Catalan originals will show no significant difference – and as a corollary nor will the frequency of the imperfect.
3. The Catalan preterite will be over-represented in translations from English as compared to translations from French – and as a corollary the imperfect will be under-represented.

The main methodological inspiration in this study is Halverson's work, but it also draws on Hareide (2017a, 2017b), who used two comparable parallel corpora to test for the impact of different patterns of salience and connectivity. Since French (like Catalan, unlike English) does have the imperfective/perfective distinction in the past, possible differences between the two parallel corpora can throw light on the influence of factors 2 and 3. The following stages can be identified in the research process:

1. Data retrieval (preterite and imperfect) with CQP from EN-CA and FR-CA (starting from the TT end) and from Catalan NTR
2. Quantification + testing for significance
3. Identification of main triggers for the preterite and the imperfect in the ST component of both parallel corpora
4. Data retrieval (simple past and other possible triggers for English, *passé simple* and *imparfait* for French) with CQP from EN-CA and FR-CA (starting from the ST end)
5. Identification of main TT matches for the ST triggers in 4
6. Determining strength of connectivity patterns between ST and TT forms
7. Hypothesis verification and refinement

Table 1 shows raw and normalised frequencies for the three sub-corpora. Results are surprising in more than one respect. Firstly, the imperfect tense is more frequent than the preterite in all three sub-corpora. This runs counter to the assumption that the preterite is the more salient of the two tenses – unless of course there is more to salience than just frequency. Secondly, the overall normalised frequency of simple past tenses (i.e. imperfect + preterite) is roughly the same in Catalan originals and translations from English (57.39 and 58.3, respectively) but remarkably lower in translations from French (46.41). That points to a less verbal style in translations from French, possibly as a result of source language tendencies. In spite of this, two of the three hypotheses formulated above (1 and 3) are confirmed, as the preterite is over-represented in EN-CA as compared both to NTR and FR-CA. Hypothesis 2, on the other hand, is not confirmed, the preterite being over-represented in FR-CA too as compared to NTR. Differences are significant in all three cases.

	CAT		EN-CAT		FR-CAT	
	Raw f	Normalised f per 1,000 words	Raw f	Normalised f per 1,000 words	Raw f	Normalised f per 1,000 words
Imperfect	52,572	33.88	39,616	29.48	14,960	26.38
Preterite (synthetic + periphrastic)	36,470	23.51	38,719	28.82	11,356	20.03

Table 1. Raw and normalised frequencies of preterite and imperfect in Catalan non-translations and in the translated components of EN-CA and FR-CA

These results only make full sense when completed with the analysis in stages 3-5 above, which will enable us to determine the strength of connectivity patterns between ST and TT forms (stage 6). Strength of connectivity is operationalised as a formula bringing together source and target concentration (Schmid 2010; see Marco 2021 for an application of this formula). Query results for English-Catalan show a higher degree of connectivity between the simple past and the Catalan preterite than between the former and the Catalan imperfect. That, together with the unquestioned salience of the English simple past in the source language network and the posited salience of the Catalan preterite in the target language network, accounts for over-representation of the preterite in EN-CA (hypotheses 1 and 3). The picture is much more complex in FR-CA. Even though methodological stages 3-6 show a high degree of connectivity between the French *passé simple* and the Catalan preterite, on the one hand, and the French *imparfait* and the Catalan imperfect, on the other, the balance between the two simple past tenses in translations from French differs significantly from the one observed in Catalan non-translations. A possible explanation for this is that the relatively high salience of the preterite in Catalan (factor 1) prevails over a high degree of connectivity between tenses across the two languages (factor 3). But more work needs to be done on the ST component of both parallel subcorpora in order to gauge the exact weight of factor 2.

References

Binnick, R.I. (1991). *Time and the Verb. A Guide to Tense & Aspect*. Oxford: Oxford University Press.

Comrie, B. (1976). *Aspect*. Cambridge: Cambridge University Press.

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*. Berlin: Mouton de Gruyter, 9-45.

Hareide, L. (2017a). The translation of formal source-language lacunas. An empirical study of the Over-representation of Target-Language Specific Features and the Unique Items Hypothesis. In M. Ji, M. Oakes, L. Defeng & L. Hareide (eds.) *Corpus Methodologies Explained. An Empirical Approach to Translation Studies*. London & New York: Routledge, 137-187.

Hareide, L. (2017b). 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 & L. Hareide (eds.) *Corpus Methodologies Explained. An Empirical Approach to Translation Studies*. London & New York: Routledge, 188-231.

Maingueneau, D. & Salvador, V. (1995). *Elements de lingüística per al discurs literari*. València: Tàndem.

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*. Berlin: Language Science Press, 27-52.

Pérez Saldanya, M. (2002). Les relacions temporals i aspectuals. In J. Solà, M.R. Lloret, J. Mascaró & M. Pérez Saldanya (dirs.) *Gramàtica del català contemporani*, vol. 3: *Sintaxi*. Barcelona: Empúries, 2567-2662.

Schmid, H.-J. (2010). Does frequency in text really instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (eds.) *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches*. Berlin: Walter de Gruyter, 101-133.

Wheeler, M.W., Yates, A., & Dols, N. (1999). *Catalan: A Comprehensive Grammar*. London & New York: Routledge.

Almost all and *presque tout*: A corpus-based study of quantity modification with English *all* and French *tout*

Jesse Marion
Université de Mons (UMONS), Belgium
jesse.marion@umons.ac.be

Whereas many have investigated the field of quantifiers (e.g. Barwise & Cooper 1981; Gärdenfors 1987; Langacker 1991, 2008; Doetjes 1997; Benninger 1999; Radden & Dirven 2007), few have ventured into that of quantity modification, i.e. the modification of quantifying expressions, as in (1-2) below. Examples (1) and (2) respectively feature the relative quantifiers *all* and *tout*, which specify a predicated quantity P in relation to a reference mass R_T which "consists [by default] of the maximal instantiation [T] of the pertinent category", e.g. *all students* in which $P = R_T$ (Langacker 1991: 82-83, 86). In (1) and (2) English *all* and French *tous* are modified by *almost* and *presque* respectively, which indicate that the boundary of the predicated mass P approximates that of the reference mass R_T but that P and R_T do not completely coincide.

(1) That's why I use him in ***almost all*** my answers. (YCCQA_uk)
(2) (…) Beaucoup font aussi la lecture de vidéo, comme VLC qui peut lire ***presque tous*** les formats vidéo existant (…) (YCCQA_fr)
'(…) Many can also read videos, such as VLC which can read ***almost all*** existing video formats (…)'

This study focuses on one type of quantification, relative quantification, and explores the possible ways in which relative quantifiers can be modified in terms of quantity. The two quantifiers under scrutiny are English *all* and French *tout* (in all its inflected forms). The contrastive set-up of the study will allow us to investigate any cross-linguistic differences in the quantity modification potential of relative quantifiers.

Building on Paradis (1997, 2000, 2001), Njende *et al.* (2015, 2017) revealed co-selection restrictions between specific modifier and quantifier types. They found that absolute quantifiers like *many* and *few* take scalar modifiers, which modify – upwards or downwards – the range expressed by the quantifier on an implied open quantification scale (e.g. *very many, rather few*). Relative quantifiers such as *all* or *none* take proportional modifiers, which compare the actually predicated quantity or mass to the reference mass, indicating whether it either approximates (e.g. *almost all/none*) or reaches it (e.g. *absolutely all/none*).

This study aims to verify Njende *et al*.'s (2015, 2017) findings for *all* and complement them with contrastive data for French *tout*. This will contribute to inventorying the different types of quantity modification the items under scrutiny allow. English *all* and French *tout* were selected for this study as they can be modified by a variety of adverbs and are relatively frequent in everyday speech.

Data were extracted from a multilingual lower-register written corpus and two spoken corpora. Standard corpora such as Collins Wordbanks Online for English and Frantext for French returned too few instances of quantity modification. As quantity modification is in many respects similar to degree modification, lower-register and spoken corpora were consulted as they have proven to be fruitful sources of data on degree modification. The first corpus consulted is the 29-million-word Yahoo-based Contrastive Corpus of Questions and Answers (YCCQA, De Smet 2009), which is available in English, French, German, and Spanish, and covers the time period 2006 to 2009 (De Smet 2009). Datasets of 250 instances were

extracted with AntConc (Anthony 2010) from both the English and French YCCQA subcorpora, provided that sufficient data was available.

The English YCCQA data are supplemented with another 250 instances taken from the British National Corpus 2014 Spoken (BNC2014, Love *et al.* 2017), accessed through SketchEngine. The BNC2014 totals 10,495,185 words from 1,251 conversations by 672 speakers between 2012 and 2016. The French YCCQA data are supplemented with spoken data from both the Corpus de Français Parisien Parlé des années 2000 (CFPP2000, Branca-Rosoff *et al.* 2012), which contains 744,159 words from circa 58 hours of audio files, and the corpus Traitement de Corpus Oraux en Français (TCOF, "Analyse et traitement"), which totals 1,542,562 words from circa 146 hours of audio files.

On the basis of the data selected, a monolingual and contrastive analysis can be performed of *all* and *tout* in informal registers in written language, and a monolingual study can be carried out for each quantifier in spoken language. Particular attention will go to the contexts in which quantity modification emerges, as well as to the (in)animate nature and semantics of the nouns quantified.

Preliminary data analysis seems to indicate that quantity modification is rarer in French than in English. This may find an explanation in the considerable number of instances of *not all N* in the English data. *Not all* typically has scope over the subject of the sentence, as in (3). The French data returned no such uses of *pas tout*. This could be linked to the position markers of negation favour in the sentence in the languages under study.

(3) ***Not all*** white people are like that. (YCCQA_uk)

Therefore, we posit that quantity modification is overall more frequent in English than in French, and that negative quantity modifiers (such as *not* or *pas*) are common in English only. In addition, preliminary results seem to suggest that animate nouns modified by one of the quantifiers under study most often serve as subjects in the sentence (as in (3) above). Quantified inanimate nouns, by contrast, act most often as direct objects, as in (4).

(4) You had ***almost all*** the advice you need by fellow Italians. (YCCQA_uk)

This study will thus allow to catalogue the quantity modifiers for *all* and *tout* in informal and spoken registers, will contribute to a better understanding of quantity modification in general and of its relation to the type of quantification involved, and will compare and contrast how *all* and *tout* in particular undergo quantity modification. More precisely this study will explore the semantics of the nouns quantified and their possible influence on the choice of quantifier, as well as the frequency to which proportional and totality modifiers are involved in the quantity modification of the relative quantifiers *all* and *tout*.

References

Analyse et traitement informatique de la langue française - UMR 7118 (ATILF) (2020). *TCOF : Traitement de Corpus Oraux en Français* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v2.1, www.hdl.handle.net/11403/tcof/v2.1.

Anthony, L. (2010). *AntConc* (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. www.antlab.sci.waseda.ac.jp.

ATILF. *Base textuelle Frantext* (En ligne). ATILF-CNRS & Université de Lorraine. 1998-2021. www.frantext.fr.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. Linguistics and Philosophy. *Linguistics and Philosophy*, 4(2), 159-219.

Benninger, C. (1999). *De la quantité aux substantifs quantificateurs*. Metz : University of Metz PhD thesis. Available online at www.gallica.bnf.fr/ark:/12148/bpt6k33703112.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., Pires, M. (2012). *Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*. www.cfpp2000.univ-paris3.fr/CFPP2000.pdf.

De Smet, H. (2009). *Yahoo-based Contrastive Corpus of Questions and Answers (YCCQA).* varieng.helsinki.fi/CoRD/corpora/YCCQA.
Doetjes, J. S. (1997). *Quantifiers and selection: On the distribution of quantifying expressions in French, Dutch and English*. The Hague: Holland Academic Graphics.
Gärdenfors, P. (Ed.) (1987). Generalized Quantifiers. *Linguistic and Logical Approches*, 31. Dordrecht: D. Reidel Publishing Company.
Langacker, R. W. (1991). *Foundations of Cognitive Grammar* (Vol. 2, Descriptive Application). Stanford, CA: Stanford University Press.
Langacker, R. W. (2008). *Cognitive Grammar – A Basic Introduction*. Oxford: Oxford University Press.
Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3): 319-344.
Njende, N. M., Davidse, K., & Ghesquière, L. (2017). Precious few and practically all: A cognitive grammar approach to the modification of quantifiers. *Leuven Working Papers in Linguistics*, 34, 96-121.
Njende, N. M., Ghesquière, L., & Davidse, K. (2015). *Re-assessing the absolute quantification restriction in existential constructions*. Paper presented at the 48*th* Annual Meeting of the Societas Linguistica Europaea (SLE 48), 2nd-5th September 2015, Leiden, the Netherlands.
Paradis, C. (1997). *Degree modifiers of adjectives in spoken British English*. Lund: Lund University Press.
Paradis, C. (2000). Reinforcing adjectives: A cognitive semantic perspective on grammaticalization. In R. Bermudez-Otero, D. Denison, R. Hogg & C. B. McCully (Eds.), *Generative theory and corpus studies* (Topics in English linguistics 31), 233- 258). Berlin: Mouton de Gruyter.
Paradis, C. (2001). Adjectives and boundedness. *Cognitive Linguistics*, 12(1), 47-65.
Radden, G., & Dirven, R. (2007). Cognitive English Grammar (Vol. 2, Cognitive Linguistics in Practice). Amsterdam: Benjamins.
Sketch Engine: www.sketchengine.eu.
Wordbanks Online: www.wordbanks.harpercollins.co.uk/.

Translating repetition: A corpus study of the translation of repeated reporting verbs in the *Harry Potter* series and its Italian version

Lorenzo Mastropierro
University of Insubria
lorenzo.mastropierro@uninsubria.it

The repetition of a linguistic feature is a reflection of its functional relevance (Mahlberg 2010: 297). In stylistic terms, repetition is one of the main devices used to create foregrounding (Leech & Short 2007). Toolan (2012: 22) explains that literary texts in particular exploit repetition in many different ways, establishing networks of "echoically linkable" structures that achieve a "depth of texture [...] atypical in non-literary discourse". There is extensive research on the significance of repetition in specific literary texts (see, for example, Prusse 2012 and Paton 2009), including corpus approaches (Mahlberg 2010, 2013) that have provided quantitative insights into the role of repeated patterns in literature. However, the existing corpus-based studies on the translation of repeated features in literary texts (Čermáková & Fárová 2010; Čermáková 2015; Mastropierro & Mahlberg 2017, Mastropierro 2020) show that translators tend to avoid repetition in favour of variation, to the detriment of the stylistic effects that repetition may be contributing to in the source text. This paper aims to shed further light on the translation of repetition, through a study of repeated reporting verbs (e.g. *Harry said*, *cried Ron*, or *whispered Hermione*) in the *Harry Potter* book series and its Italian translation. This study's objective is to identify potential factors that could prompt translators to translate the same verbs in multiple different ways, with the risk of altering the role that repetition plays as a stylistically relevant feature.

This paper builds on a previous study (Mastropierro 2020) in which I explore the translation of reporting verbs in the Italian versions of J. K. Rowling's seven *Harry Potter* books. The findings of this study show that both translators (Marina Astrologo translated the first two novels, Beatrice Masini translated the remaining five) use a much wider lexical variety to translate the source text reporting verbs. This variety and the change of reporting verb category (see Caldas-Coulthard 1987) are shown to have potential consequences for character development throughout the series. That is, the balance between the different categories of verbs is altered in translation, affecting the potential of reporting verbs to reflect the changes the characters go through between the beginning and the end of the series. In this paper, I build on these findings and investigate whether a relationship exists between frequency, category, and meaning of the repeated verbs and the extent to which they are translated in multiple ways. To do so, I use the *InterCorp* parallel corpus (Čermák & Rosen 2012) to identify reporting verbs associated with the three central protagonists of the seven novels (Harry, Ron, and Hermione), and their translations in the Italian books. CQL queries are used to retrieve all verbs occurring alongside character names following direct speech (e.g. *"[…]" Harry said*, *"[…]" urged Hermione*). The lists are manually checked to remove all verbs that are not reporting verbs, and parallel concordance lines are then employed to identify the translations of the remaining reporting verbs. Once all English verbs and their Italian translations have been identified for each character, I investigate potential relations between frequency, category, and meaning of the repeated verbs, and in how many different ways they are translated.

To start with, this study explores whether a correlation exists between the frequency of the repeated verb and the number of its different translations. It may be expected that the number of different translations is proportional to the frequency of the original verb, but preliminary results show that this is not always the case. For example, *said* occurs 372 times in the first two books in relation to Harry and is translated in 41 different ways in the target texts. Muttered *occurs* 15 times and is translated in 6 different ways, while *gasped* occurs 8 times and has 4 different Italian translations. However, *asked* occurs 28 times but has no multiple translations. The relationship between verb frequency and translation variations therefore

needs to be verified; taking into account all of the reporting verbs, I test whether or not a recognisable correlation exists between the frequency of the source text item and the number of different translations. Building on this quantitative perspective, this paper then investigates more qualitatively whether meaning and category of verb influence its translation into multiple Italian variants or not. It is assessed how many different meanings and different translations each English verb has, as well as the category the verb belongs to, using Caldas-Coulthard's (1987) taxonomy. In this way, it is possible to explore the question of whether factors like polysemy of the source text item, existence of multiple translational equivalents in the target language, or verb type (for instance, neutral vs. metapropositional vs. paralinguistic, etc.) are related to the choice of translating the same repeated item in different ways.

By combining quantitative and qualitative dimensions, this paper offers important insights into the translation of repetition. A more detailed description of the translation of repeated items can provide better understanding of the phenomenon in the context of professional practice, with important implications for 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 of the original.

References

Caldas-Coulthard, C. R. (1987). Reported speech in written narrative texts. In M. Coulthard (ed.), *Discussing Discourse*. Birmingham: University of Birmingham, 149-167.

Čermáková, A. (2015). Repetition in John Irving's novel *A Widow for One Year*: A corpus stylistics approach to literary translation. *International Journal of Corpus Linguistics 20*(3), 355-377.

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

Čermáková, A. & Fárová, L. (2010). Keywords in *Harry Potter* and their Czech and Finnish translation equivalents. In F. Čermák, P. Corness & A. Klégr (eds.), *InterCorp: Exploring a Multilingual Corpus*. Praha: NLN, 177-188.

Leech, G. & Short, M. (2007). *Style in Fiction. A Linguistic Introduction to English Fictional Prose (2nd ed.)*. Harlow: Pearson Longman.

Mahlberg, M. (2010). Corpus linguistics and the study of nineteenth-century fiction. *Journal of Victorian Culture 15*(2), 292-298.

Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. London: Routledge.

Mastropierro, L. (2020). The translation of reporting verbs in Italian: The case of the *Harry Potter* series. *International Journal of Corpus Linguistics 25*(3), 241-269.

Mastropierro, L. & Mahlberg, M. (2017). Key words and translated cohesion in Lovecraft's *At the Mountains of Madness* and one of its Italian translations. *English Text Construction 10*(1), 78-105.

Paton, S. (2009). Time-Lessness, simultaneity and successivity: Repetition in Beckett's short prose. *Language and Literature 18*(4), 357-366.

Prusse, M. (2012). Repetition, difference and chiasmus in John McGahern's narratives. *Language and Literature 41*(4), 363-380.

Toolan, M. (2012). Poems: Wonderfully repetitive. In R. Jones (ed.), *Discourse and Creativity*. London: Pearson, 17-34.

A POS-gram study of case relations in Serbian, Italian and English: The role of articles, inflection and word order

Maja Miličević Petrović[1], Dragana Radojević[2]
University of Bologna[1], University of Belgrade[2]
maja.milicevic2@unibo.it[1], dragana.radojevic@fil.bg.ac.rs[2]

Since the advent of corpus approaches in contrastive language studies, a range of phenomena have been explored through different types of multi-word combinations, recognised as key building blocks of languages in terms of both meaning and structure. One prominent kind have been *n*-grams – uninterrupted sequences of *n* words (e.g., *at the end of* in English). The focus has mostly been on contrasting *n*-grams that encode specific semantic categories or discourse functions (e.g., place expressions in Čermáková and Chlumská 2017, or metadiscursive bundles in Granger 2014). Somewhat less attention has been given to structural phenomena (but see Lazić 2017 for a comparison of structurally classified *n*-grams in English and Serbian).

A related type of combinations are POS-grams – uninterrupted sequences of parts of speech (e.g., article-noun), possibly combined with other grammatical categories (e.g., definite article-plural noun). POS-grams carry morphosyntactic information and they can reveal important facts about languages, both at a general level (e.g., concerning word order preferences), and in relation to specific phenomena (see Chlumská 2018 on English-influenced POS-grams typical of translations into Czech).

With both *n*-grams and POS-grams, multiple methodological issues affect contrastive studies. First, selecting appropriate lengths of *n*/POS-grams to compare is not straightforward, especially when studying analytical vs. inflectional languages: the English four-gram *at the end of*, for instance, has a bigram equivalent *na kraju* 'at end' in Serbian; this is why authors such as Ebeling & Ebeling (2013) and Granger (2014) suggested not limiting contrastive studies to single-length bundles. In addition, cross-linguistic comparisons are known to be affected by word order, and fixed-order languages like English tend to have more frequent recurrent patterns than free-order languages (e.g., Slavic; Chlumská & Lukeš 2018).

With this background in mind, we use POS-grams to study the expression of oblique case relations in Serbian, Italian and English. Italian and English are predominantly analytical in the nominal domain, expressing case relations via prepositional phrases (e.g. *della bambina – of the girl*); an exception is given by the Saxon genitive in English (*girl's*), and by cases primarily determined through word order – the accusative and (in English) partly dative; the latter will be disregarded in this study due to the difficulty of automatic retrieval. The two languages differ in articles, which in Italian have more forms and can sometimes be merged with prepositions (e.g., *di* 'of' + *la/le* 'the' > *della/delle* 'of the (fem.sg/pl.)'). Serbian, on the other hand, is an article-less markedly inflectional language, with oblique cases morphologically encoded on nouns (e.g., for *kraj* 'end': genitive *kraj**a***, dative/locative *kraj**u***, instrumental *kraj**em***). Serbian also uses prepositional phrases, in which the case of the noun is determined by the preposition (***na*** *kraj**u*** 'at end', ***do*** *kraj**a*** 'until end').

The research question we address is whether Serbian, Italian and English, once their structural differences are taken into account, display similar patterns of oblique case use. We adopt a corpus-driven approach, and our additional objective is to test the POS-gram method on a three-way comparison between typologically distinct languages. Our analysis relies on comparable corpora from the Web-as-Corpus family: srWaC (Ljubešić & Klubička 2014), itWaC and ukWaC (Baroni et al. 2009). These corpora were created following the same general principles, they contain a variety of text genres, and are very large (the smallest, srWaC, counting half a billion words).

We accessed the corpora via the Sketch Engine platform (Kilgarriff et al. 2014). We first compared preposition-noun combinations, as typical expressions of oblique case relations in all three languages. POS-bigrams and POS-trigrams were extracted. For Italian, we looked at bare preposition + noun sequences (capturing bigrams such as *a casa* 'at home'), prepositions with attached articles + nouns (*della bambina*), and bare prepositions + articles + nouns (*con la bambina* 'with the girl'). The first and third types were obtained for English, while only the first one was extracted for Serbian (with all oblique cases joined together; *kod kuće* 'at home', *sa devojčicom* 'with girl'). The respective frequencies are shown in Table 1, under [1].

	[1] preposition + noun	[2] preposition + (adjective) + noun	[3] (preposition) + noun
srWaC	47,415	62,030	151,187
itWaC	94,436	101,813	101,813
ukWaC	63,203	77,829	80,859

Table 1. Normalised frequencies (per million words) of the studied POS-grams.

To capture word order differences, focusing on adjective placement as their most immediate source in the nominal domain (Italian largely having post-nominal, and Serbian and English pre-nominal modifiers), we also included preposition + (article) + adjective + noun POS-grams, and added their frequencies to the initial frequency counts ([2] in Table 1). Finally, we proceeded to add to the counts (see [3] in Table 1) preposition-less instances of genitive, dative and instrumental cases in Serbian (exemplified by unigrams such as *devojčici* 'to girl'), as well as the Saxon genitive in English (*girl's*), as such POS-unigrams directly correspond to Italian prepositional POS-bigrams/trigrams.

The results point to some dissimilarities between languages. Preposition-noun sequences are much more frequent in Italian than in Serbian and English, even when pre-nominal adjectives are taken into account. However, Serbian takes the lead if oblique-case POS-unigrams are also considered. A closer inspection reveals that this change is primarily due to the high frequency of preposition-less genitive forms in Serbian (74,941 pmw), also used in some situations that would not be treated as oblique in Italian and English (e.g., in temporal expressions such as *ove godine* 'this year', or following numerals, as in *tri godine* 'three years'). We are currently conducting a set of additional comparisons to address such functional differences. We will also complete the analysis with frequency estimates of preposition-less datives in English.

Judging from our study, cross-linguistic differences in articles can be overcome with careful POS-gram selection, while inflection and word order discrepancies do hinder automatic comparisons, requiring additional analyses. Even though our results indicate that inflection affects the results more than word order, we will also discuss the possibility of using POS-skip- grams to capture word order differences in a more principled way (similarly to what was suggested by Chlumská & Lukeš 2018 for Czech).

References

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3), 209-226.

Chlumská, L. (2018). Prominent POS-grams and n-grams in translated Czech in the mirror of the English source texts. In M. Fidler & V. Cvrček (Eds), *Taming the Corpus: From Inflection and Lexis to Interpretation.* Cham: Springer, 99- 118.

Chlumská, L. & Lukeš, D. (2018). Comparing the incomparable? Rethinking n-grams for free word-order languages. Presentation at UCCTS 2018, Louvain-la-Neuve, 12-14 September 2018.

Čermáková, A. & Chlumská, L. (2017). Expressing PLACE in children's literature: Testing the limits of the n-gram method in contrastive linguistics. In T. Egan & H. Dirdal (Eds), *Cross-linguistic Correspondences: From Lexis to Genre.* Amsterdam: John Benjamins, 75- 95.

Ebeling, J. & Ebeling, S. O. (2013). *Patterns in Contrast*. Amsterdam: John Benjamins.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14(1), 58-72.

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. http://www.sketchengine.eu/

Lazić, K. (2017). *Učestali leksički spojevi u engleskom jeziku biotehničke struke: Korpusna analiza radova izvornih i neizvornih govornika*. PhD dissertation, University of Belgrade.

Ljubešić, N. & Klubička, F. (2014). {bs,hr,sr}WaC - Web corpora of Bosnian, Croatian and Serbian. In F. Bildhauer & R. Schäfer (Eds), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*. Gothenburg: Association for Computational Linguistics, 29-35.

The imperfective/perfective aspect in Catalan and its German triggers

Teresa Molés-Cases[1]
Universitat Politècnica de València[1]
temoca1@upv.es[1]

The inspiration for this contribution arises from the observation that Catalan and German diverge with regard to aspect. More specifically, whereas the imperfective/perfective aspect distinction is morphologically marked in the past in the Catalan verb system (Alturo Monné 2008: 14), German verb forms have no morphological means to express aspect (Heinold 2015). Catalan distinguishes between the imperfect and the preterite (e.g. *Mentre parlava amb ell, em vaig adonar* […] "While I was talking to him, I realised that…", Pérez Saldanya 2002: 2578), whereas German only has one simple past tense (e.g. *Peter war gerade dabei zu essen, als Maria anrief* "Peter was just about to eat when Maria called", Heinold 2015: 11). This makes German significantly different from other Germanic languages which still express some aspectual contrasts through verb forms (e.g. English) (González and Diaubalick 2020: 310) (e.g. *John was reading when I entered*, Comrie 1976: 3). However, in German it is possible to express aspectual nuances by other means (Comrie 1976: 8). For instance, as indicated by Heinold (2015: 64), the adverb *gerade* "just" can express progressive aspect, whereas the adverbs *immer* "always" and *gewöhnlich* "usually" denote habitual aspect (see Comrie 1976: 25, for a classification of aspectual oppositions). The German counterparts of the Catalan imperfect and preterite are the *Präteritum* (also called *Schreibtempus für literarische Texte* "writing tense for literary texts", Dreyer and Schmitt 2000: 325) (e.g. *Gestern war ich zu Hause* "Yesterday I was at home") and the *Perfekt* (e.g. *Sie haben Wasser getrunken* "They have drunk water"), which is predominant in spoken situations (Kessel and Reimann 2017: 94). Although there is some controversy in the literature about the differences between these two tenses (see Schwenk 2012, for an overall discussion on past verb tenses in German), recent contributions have concluded that these are mainly stylistic in nature (i.e. formal/informal, written/spoken, Heinold 2015: 102–111). These contrasts between the Catalan and the German verb systems constitute a translation problem, as German-Catalan translators have to mark aspect explicitly when referring to past events (Ainaud *et al*. 2020: 172; Lawick 2009: 198). Moreover, they can lead to the existence of several published versions of literary texts including different past tenses (Haßler 2016: 273–298, on contrastive issues between Romance and Germanic languages with this regard).

This contribution derives from an ongoing research project which aims to test out the Gravitational Pull Hypothesis (GPH) (Halverson 2003, 2017) on a number of indicators (Marco 2021a), including the imperfective/perfective aspect distinction in Catalan (Marco 2021b). Halverson's (2003: 197) GPH highlights the importance of considering cognitive salience and asymmetry in the semantic structure when studying translation universals or patterns. Halverson (2017: 14) identifies three potential causes, or factors, of translational effects in the translation process: 1) source language salience (gravitational pull), 2) target language salience (magnetism) and 3) link strength effects (connectivity between the source and the target languages). Factors 1 and 2 lead to overrepresentation, factor 3 may lead to over- or underrepresentation, depending on the strength of the connectivity patterns. For the purposes of this study, some sections of the COVALT corpus (Valencian Corpus of Translated Literature, Universitat Jaume I, Spain) will be examined. COVALT was originally created as a multilingual parallel corpus containing complete narrative works originally written in English, French and German with their Catalan translations published in the autonomous region of Valencia between 1990 and 2000. The original corpus has since been extended to include a comparable component of Catalan non-translations too. In the first stage of this research study the following subsections will be examined: a comparable Catalan subcorpus of original narrative works (1,551,521 tokens), a collection of narrative texts originally written in German (546,178 tokens) and the corresponding translations into Catalan (604,966 tokens). The corpus, which is

indexed in IMS Open Corpus WorkBench, has been analysed using the Corpus Query Processor (web version) (CQPweb).

Table 1 details the stages of the present contribution:

Stage	Corpora	Corpus search
1. Data retrieval, quantification and significance test	a) Translated texts into Catalan (from German) b) Original texts in Catalan	Forms of imperfect and preterite
	c) Original texts in German	German triggers (see Table 3)
2. Determining strength of connectivity patterns between German and Catalan forms		

Table 2. Stages of the research study.

In this study two hypotheses are posited: a) the Catalan preterite is overrepresented in translations from German in comparison with Catalan original texts, and b) the Catalan imperfect is underrepresented in translations from German as compared to Catalan original texts. The rationale for these hypotheses stems from the observation that the preterite is more salient than the imperfect in narrative scenarios (Maingueneau and Salvador 1995: 42; Pérez Saldanya 2002: 2579).

Next the data available so far will be presented. Table 2 indicates the frequency of imperfect and preterite forms in Catalan non-translated (CAT) and translated texts (DE-CAT):

	CAT		DE-CAT	
Catalan verb tense	Raw frequency	Normalised frequency per 1,000 words	Raw frequency	Normalised frequency per 1,000 words
Imperfect	52,572	33.88	17,406	28.77
Preterite (synthetic + periphrastic)	35,159	22.66	19,652	32.48

Table 2. Frequencies of imperfect and preterite forms in Catalan non-translated and translated texts.

In texts originally written in Catalan (CAT), the imperfect appears to be more frequent than the preterite. This may be said to contradict the assumption made above that the preterite is more salient than the imperfect – unless it is also assumed that there is more to salience than just frequency. Be that as it may, these results are in marked contrast to those yielded by the translated component (DE-CAT), where preterite clearly outnumber imperfect forms. This seems to confirm the two hypotheses posited above.

Table 3 gives a full account of the analysis performed on the German-Catalan parallel corpus starting from the target end. A 10 % thinning was applied to the overall results yielded by the query and 3706 bilingual concordances were analysed with a view to identifying the major German triggers of the imperfect and preterite forms.

As can be observed, the *Präteritum* is the most common German trigger of both the imperfect and preterite forms in the Catalan translated texts. This is not surprising, since the *Präteritum* is the most common verb tense in written narrations. The next steps are to retrieve the forms of *Präteritum* from the German source text component to determine the strength of connectivity patterns between the German *Präteritum* and the Catalan imperfect and preterite forms (see Table 1), as the picture offered by Table 3 is partial, analysis having proceeded from the target end only. At a later stage of the research project, the data yielded by the German-Catalan subcorpus will be compared with data from the French-Catalan

subcorpus, as translation solutions may reflect the different aspectual configurations of simple past tenses in German and French.

Triggers	Imperfect (DE-CAT)	%	Preterite (synthetic + periphrastic) (DE-CAT)	%
Präteritum	1250	71.80	1782	90.69
Other (nominalizations, *weiter*, etc.)	101	5.80	49	2.49
Konjunktiv I	78	4.48	9	0.46
Plusquamperfekt	30	1.72	30	1.53
Konjunktiv II	30	1.72	11	0.56
Präsens	29	1.67	12	0.61
Perfekt	28	1.61	43	2.19
Noise	195	11.20	29	1.48
Total	1741	100.00	1965	100.00

Table 3. The German triggers found in the Catalan translated texts.

References

Ainaud J., Espunya, A. & Pujol, D. (2020). *Manual de traducció anglès-català*. Barcelona: Universitat Pompeu Fabra.

Alturo Monné, N. (2008). *La semàntica verbal del català: la representació dels esdeveniments*. Doctoral Thesis: Universitat de Barcelona.

Comrie, B. (1976). *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.

Dreyer, H. & Schmitt, R. (2000). *Lehr- und Übungsbuch der deutschen Grammatik*. Ismaning. Hueber.

González, P. & Diaubalick, T. (2020): Subtle Differences, but Rigorous Implications: German and Dutch Representation of Tense-aspect Features. Evidence from SLA. In G. De Vogelaer, D. Koster & T. Leuschner (eds.) *German and Dutch in Contrast: Synchronic, Diachronic and Psycholinguistic Perspectives*. Berlin: De Gruyter, 299-328.

Halverson, S. (2003). The Cognitive Basis of Translation Universals. *Target* 15(2), 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. New Methodological and Theoretical Traditions.* Berlin: Mouton de Gruyter 2017, 9-46.

Haßler, G. (2016). *Temporalität, Aspektualität und Modalität in romanischen Sprachen*. Berlin: De Gruyter.

Heinold, S. (2015). *Tempus, Modus und Aspekt im Deutschen*. Tübingen: Narr Verlag.

Kessel, K. & Reimann, S. (2017). *Basiswissen Deutsche Gegenwartssprache*. Tübingen: A Francke Verlag.

Lawick, H. van (2009): *Manual de traducció alemany-català*. Vic: Eumo.

Maingueneau, D. & Salvador, V. (1995). *Elements de lingüística per al discurs literari*. València: Tàndem.

Marco, J. (2021a). 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*. Berlin: Language Science Press, 27-52.

Marco, J. (2021b). *The Gravitational Pull Hypothesis and Imperfective/Perfective Aspect in Catalan Translated and Non-translated Literary Texts*. Unpublished manuscript.

Pérez Saldanya, M. (2002). Les relacions temporals i aspectuals. In J. Solà, M.R. Lloret, J. Mascaró & M. Pérez Saldanya (eds.) *Gramàtica del català contemporani*, vol. 3: Sintaxi. Barcelona: Empúries, 2567-2662.

Schwenk, H. J. (2012). Die Vergangenheitstempora im Deutschen und ihr semantisches Potential. *Lublin Studies in Modern Languages and Literature* 36, 35-49.

The Subtitling of Taboo Language Terms in the French version of *Orange Is the New* Black: A Corpus-Based Analysis

Eponine Moreau
American Studies Center, Umons
eponine.moreau2@umons.ac.be

In recent years, what Bucaria (2007) calls the "massive importation of audiovisual products, mainly from English speaking countries" has resulted in the emergence of new prospects in the field of audiovisual translation research. One of the main features of these audiovisual products is the high frequency of taboo language words used in the dialogues. The translation of these terms can be challenging for the audiovisual translators thus making it an interesting subject for an increasing number of researchers.

This paper reports on a study into the strategies used to translate taboo terms in the French subtitled version (FST) of the first season of *Orange Is the New Black* (*OITNB*). Adopting a corpus-based approach to analyse the subtitling of specific English swear words and their variants into French, this study aims to shed some light on the translation of offensive and taboo language in audiovisual programmes.

In the remainder of this abstract, we will briefly describe the corpus and present the research questions that this study addresses together with the methods used to answer them.

The Selection of the Corpus and Data Collection

OITNB is a US dramedy series that was first released on the streaming service Netflix in July 2013. This corpus was selected for two main reasons. First, as a Netflix original series, *OITNB* is one of the TV shows that has contributed to the emergence of binge-watching (Matrix 2014: 119). As such, the series could also be the subject of further research in the field of reception studies. Second, the TV show deals with many controversial themes which, together with the setting of the series itself (a women's federal prison), guaranteed that the dialogues would include many offensive/taboo terms. A preliminary quantitative analysis has indeed shown that, throughout the first season of *OITNB*, the words *shit* and *fuck* (and their variants) were uttered 246 and 430 times respectively.

The whole corpus consists of 148037 words. The number of words and subtitles for each episode and each version is presented in the following table:

Word Count/Episode	English OV	FST	#Subtitles	Episode Length
Episode 1	5183	4878	811	0:52:15
Episode 2	3910	3756	678	0:53:00
Episode 3	6180	5685	936	0:57:31
Episode 4	5295	5044	831	0:55:20
Episode 5	6340	6148	943	0:55:28
Episode 6	5733	5657	832	0:57:29
Episode 7	5510	5231	816	0:57:53
Episode 8	5589	5523	827	0:59:39
Episode 9	6168	5843	945	0:59:47
Episode 10	6584	6112	974	0:54:35
Episode 11	6567	6287	987	01:00:03
Episode 12	6790	6599	963	01:01:00
Episode 13	5854	5571	853	01:01:13
Total	75703	72334	11396	12:25:13

To avoid the problems associated with using online transcripts, this *ad-hoc* parallel corpus was edited and aligned manually thanks to .xml files downloaded from Netflix and converted into Excel sheets.

Research Questions and Methods

The analysis of the corpus aims to answer the following research questions:

- *What are the most frequent offensive/taboo categories in the English sub-corpus and in the French sub-corpus respectively*?

The taboo terms retrieved in the corpus will be divided into the same categories as the ones used by Soler Pardo who relied on the work of Timothy Jay and McEnery to classify the occurrences under the following types: sex-related insults, excrement and human waste insults, body parts, religious insults, incest, prostitution, racial slurs, cross-categorised insults/ swear words, physical or mental disability, bodily functions, animal-related insults, and homophobic insults. A quantitative analysis of the categories will be carried out using the software Sketch Engine. It will also be interesting to compare the results yielded from both sub-corpora.

- *What are the effects achieved by the translation strategies used to render the most frequent taboo terms (the f-word and the s-word) and their variants in the FST?*

Drawing on both Bucaria's and Soler Pardo's methodology, I will provide a fairly literal back-translation of each French subtitle under scrutiny to determine and classify the effects achieved by the translation strategies. Bucaria identifies four approaches used to deal with taboo language words:

- Complete omission: the potentially offensive/taboo word has been completely deleted or replaced with a totally neutral comment.
- Weakening: the potentially offensive/taboo word has been translated by a harmless or less vulgar equivalent.
- Close rendering: the potentially offensive/taboo word has been translated literally or by an equivalent that conveys the same message.
- Increased effect: cases in which the pragmatic intensity of the potentially offensive/taboo word has been increased in the target version.

A quantitative analysis of these approaches should then allow me to determine how much of the offensive/taboo load has been transferred to FST.

- *In case of complete omission, can the strategy used always be justified by potential spatio-temporal constraints?*

To address this question, I will use the same table of equivalence between time and space for a reading speed of 180 words per minute that Ávila-Cabrera (2016) adapted from Díaz Cintas and Remael (2014).

The calculations presented in the table are based on WinCAPS, a professional subtitling software. They stipulate the maximum number of characters per subtitle according to its duration.

180 words per minute	Seconds: frames	Spaces	Seconds: frames	Spaces

		01:00	17	02:00	35
		01:04	20	02:04	37
		01:08	23	02:08	39
		01:12	26	02:12	43
		01:16	28	02:16	45
		01:20	30	02:20	49
Seconds: frames	Spaces	Seconds: frames	Spaces	Seconds: frames	Spaces
03:00	53	04:00	70	05:00	78
03:04	55	04:04	73	05:04	78
03:08	57	04:08	76	05:08	78
03:12	63	04:12	76	05:12	78
03:16	65	04:16	77	05:16	78
03:20	68	04:20	77	05:20	78

Every instance where the offensive terms have been omitted will be verified to determine whether the omission is the result of technical constraints, i.e. would the subtitlers have had enough spaces left to choose a strategy that would have enabled them to avoid the complete omission of the taboo term(s)? The guidelines provided by Netflix will also be taken into account to determine whether the omission might have been preferred to meet the company's style requirements in terms of subtitles.

Preliminary results

The preliminary analyses carried out on the English and French sub-corpora have shown that the most frequent offensive/taboo categories are the sex-related insults and the excrement and human waste insults. This is the reason why the second question will be addressed through the quantitative and descriptive analyses carried out on the translation strategies used to render the most frequent four-letter words and their variants in the first season of *OITNB*. Preliminary results have shown that the decision to omit them in the FST is not always motivated by spatio-temporal constraints. The research questions addressed in this paper could help determine whether there has been any type of text manipulation in the subtitles of the first season of *OITNB*.

References

Ávila Cabrera, J. J. (2016). The treatment of offensive and taboo terms in the subtitling of Reservoir Dogs into Spanish. *TRANS. Revista de Traductología*, 20, 24–40. https://doi.org/10.24310/trans.2016.v0i20.3145

Beseghi, M. (2016). WTF! Taboo Language in TV Series: An Analysis of Professional and Amateur Translation. *Altre Modernità*, 215-231. https://doi.org/10.13130/2035-7680/6859.

Bednarek, M. (2019). "Don"t say crap. Don't use swear words.' – Negotiating the use of swear/taboo words in the narrative mass media. Discourse, Context & Media, 29, 100293. doi:10.1016/j.dcm.2019.02.002

Bucaria, C. (2013). Humour and other catastrophes: dealing with the translation of mixed-genre TV Series. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 0(6). https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/190

Cintas, J.D., &Remael, A. (2014). *Audiovisual translation: Subtitling.* Routledge.

Díaz-Pérez, F. J. (2020). Translating swear words from English into Galician in film subtitles: A corpus-based study. *Babel*, 66(3), 393-419.

Dewaele, J. M. (2004). The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of multilingual and multicultural development*, 25(2-3), 204-222.

Gambier, Y. (2002). Les censures dans la traduction audiovisuelle. *TTR: traduction, terminologie, rédaction, 15*(2), 203-221.

Hughes, G. (2006). *An encyclopedia of swearing: The social history of oaths, profanity, foul language, and ethnic slurs in the English-speaking world.* ME Sharpe.

Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives on psychological science, 4*(2), 153-161.

Jenner, M. (2017). Binge-watching: Video-on-demand, quality TV and mainstreaming fandom. *International journal of cultural studies, 20*(3), 304-320.

Lebtahi, Y. (2004). Télévision: Les artefacts de la traduction-adaptation: le cas de la sitcom. *Meta: journal des traducteurs/Meta: Translators' Journal, 49*(2), 401-409.

Ljung, M. (2010). *Swearing: A cross-cultural linguistic study.* Springer.

McEnery, T. (2004). *Swearing in English: Bad language, purity and power from 1586 to the present* (Vol. 1). Routledge.
Merikivi, J., Mäntymäki, M., Salovaara, A., & Zhang, L. (2016). Binge watching television shows: Conceptualization and measurement.
Pedersen, J. (2018). From old tricks to Netflix: How local are interlingual subtitling norms for streamed television?. *Journal of Audiovisual Translation,* 1(1), 81-100.
Scandura, G. (2004). Sex, lies and TV: Censorship and subtitling. *Meta: Journal des traducteurs/Meta: Translators' Journal*, 49(1), 125-134.
Sidneyeve, M. (2014). The Netflix Effect: Teens, Binge Watching, and On-Demand Digital Media Trends. *Jeunesse: Young people, texts, culture*, (6), 1.
Pardo, B. S. (2013). Translating and dubbing verbal violence in reservoir dogs. Censorship in the linguistic transference of Quentin Tarantino's (swear) words. *The Journal of Specialized Translation*, 20.
Soler-Pardo, B. (2015). *On the Translation of Swearing Into Spanish: Quentin Tarantino from Reservoir Dogs to Inglourious Basterds.* Cambridge Scholars Publishing.
Vinay, J. P., & Darbelnet, J. (2000). A methodology for translation. *The translation studies reader*, 84-93.
Wajnryb, R. (2005). *Expletive deleted: A good look at bad language.* Simon and Schuster.

Filmography

Kohan, J. (Creator and Executive Producer). (2013). *Orange Is the New Black- Season 1* [TV series]. Lionsgate Television and Netflix.

Italian Eurolect variants from a contrastive perspective. Monitoring the interlinguistic distance with English Eurolect: an approach based on Natural Language Processing methods

Laura Mori[1], Giulia Venturi[2]
Università degli Studi Internazionali di Roma - UNINT[1], Istituto di Linguistica Computazionale "A. Zampolli" – ILC-CNR - Pisa[2]
laura.mori@unint.eu, giulia.venturi@ilc.cnr.it

1. Background

The UE context is a locus of language contact where translation-mediated cross-linguistic influence clearly manifests itself through lexical variants, morphological preferences and morph-syntactic structures driven by the multilingual drafting of EU legislation. Therefore, Eurolects (EU legislative varieties) develop specific features and, at the same time, they are featured by cross-linguistically converging patterns (see Mori & Szmrecsanyi, 2020). Translation-generated linguistic similarities across Eurolects may be detected using corpora and these results may empower contrastive linguistics studies.

2. Research framework

From the variational patterns resulted from *Eurolect Observatory Multilingual Corpus* (EOMC), we set out to explore similarities between Italian Eurolect (=ItEU) and English Eurolect (=EnEU) considered the role the latter plays within "intertextual chain of documents" (Koskinen, 2008: 125) where it is "the pseudo-language for most of the Union's translations" (Felici 2015: 124).

In this paper we contrasted ItEU variants from previous corpus analysis (Mori, 2018, 2019a, 2019b, 2020a, 2020b and Mori & Venturi, in preparation) with EnEU (Sandrelli, 2018) focusing on comparable corpora of Italian Eurolect and English Eurolect as opposed to corpora of both national legislative varieties.

The analysis was carried out exploiting a linguistic profiling methodology based on Natural Language Processing tools (Montemagni, 2013). It starts from the theoretical assumption that "large numbers of counts of linguistic features are used as a text profile, which can then be compared to average profiles for groups of texts" (van Halteren, 2004) and it consists in the extraction of a wide number of linguistic phenomena (lexical, morpho-syntactic and syntactic) from automatically annotated corpora. The methodology was proven to be effective in several scenarios, demonstrating the high discriminative power to monitor social and contextual language variability (Montemagni, 2013; Brunato and Venturi, 2014).

3. Methodology

For the specific purpose of this study, the linguistic profiling of Italian and English corpora was carried out using Profiling-UD (Brunato et al., 2020), a web-based tool inspired by the methodology described above and specifically devised to be multilingual since it is based on the Universal Dependencies (UD) framework (Nivre, 2015)[9] a framework for cross-linguistically consistent treebank annotation. Profiling-UD implements a two-stage process: automatic linguistic annotation and linguistic profiling. The first step is automatically carried out by UDPipe (Straka et al., 2016), a state-of-the-art pipeline available for nearly all treebanks of UD (Zeman et al., 2019). In the second step, a set of formal properties is extracted from the different levels of linguistic annotation.

[9] https://universaldependencies.org/

The methodology was applied on the Corpus A and C of the *Eurolect Observatory Multilingual Corpus* compiled as follows:

- Corpus A: 660 directives in Italian (3.469.078 tokens) and English (3.700.533 tokens)
- Corpus C: national legislative instruments in Italian (299 texts, 2.749.725 tokens) and English (1.429 texts, 8.143.964 tokens).

4. Research objective

Our aim is to see if variation related to the ItEU is somehow convergent with EnEU. Therefore, for both languages we compared the different distribution of selected features across corpora assuming both an inter-linguistic and intra-linguistic perspective. In the first case, we compared the corpora representative of the European variety of legal language (corpus A It vs. En), while the second comparison was meant to detect similarities/differences of European corpora with respect to the national legislative varieties (corpus A It vs. C It and corpus A En vs. C En). In order to prevent a possible effect size due to the different dimensions of corpora in the two languages, we carried out all the comparisons considering corpora of the same size[10]. In addition, we compared the obtained results with the same features extracted from reference corpora here considered as representative of the ordinary language, i.e. the English Web Treebank section of the English Universal Dependency Treebank (Silveira et al. 2014) and the ISDT ("Italian Stanford Dependency Treebank") section of the Italian Universal Dependency Treebank (Bosco et al. 2013). This is meant to prevent that the differences we found were due typological differences between the two considered languages. The statistical significant difference of all comparisons has been tested using the Wilcoxon rank-sum test computed with respect to the frequency distributions of the monitored linguistic features.

5. Preliminary results and discussion

In what follows, we would like to focus on three variational trends resulting from our analysis (all resulting statistically significant, p-value < 0.001):

a) Feature related with legal language patterns (i.e. use of **passive nominal subjects**);
b) Typical Italian Eurolect pattern (i.e. **simple present** to convey the deontic modality);
c) Convergence with English Eurolect (i.e. distribution of **pre-verbal nominal subjects, post-verbal nominal subjects** and preference for the **indicative mood**).

Our preliminary results show the application of an NLP-based linguistic monitoring approach to describe variational patterns in a contrastive (intralingual and interlingual) perspective. Though we focused on Italian and English, it is worth to note that translation-induced convergence within the EU framework could be also referred to influence of French Eurolect.

6. Conclusion and further research

So far the study shows cues of convergence between Eurolects due to the translation process at work within the EU context and highlights the complexity of cross-language outcomes at play including the role of French Eurolect (Mori, 2019c).

Further research could show the heuristic potential of this methodology for attaining higher-order generalizations about language use and sociolinguistic dynamics in a linguistically superdiverse world (Vertovec, 2007).

References

Bosco, C., Montemagni, S. & Simi, M. (2013). 'Converting Italian treebanks: Towards an italian stanford dependency treebank'. In A. Pareja-Lora, M. Liakata, & S. Dipper, (eds.) *7th Linguistic Annotation Workshop and Interoperability with Discourse*, 61-69.

[10] Namely, we used all corpora maintaining their original composition but we reduced the size of the English Corpus C to 607 texts for a total of 2.703.788 words, in order to make it comparable to the Italian Corpus C.

Brunato, D., Cimino A., Dell'Orletta F., Montemagni S., Venturi G. (2020). Profiling-UD: a Tool for Linguistic Profiling of Texts. *Proceedings of 12th Edition of International Conference on Language Resources and Evaluation* (LREC 2020). Marseille, France, 11-16 May, 2020.

Brunato, D. & Venturi, G. (2014). Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici. In D. Tiscornia, F. Romano, M.T. Sagri (a cura di), *Diritto, Linguaggio e tecnologie dell'informazione*. *Informatica e Diritto*, 1, 111-142.

Felici, A. (2015). Translating EU legislation from a lingua franca: Advantages and disadvantages'. In S. Šarcevic (Ed.), *Language and culture in EU law: Multidisciplinary perspectives*, London: Ashgate, 123–140.

Koskinen, K. (2008) *Translating Institutions: an Ethnographic Study of EU Translation*, New York: St. Jerome Publications.

Montemagni, S. (2013). Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, XLII(1),145-172.

Mori, L. & Venturi, G. (under preparation). Leggi italiane di derivazione europea tra direttive dell'UE e leggi nazionali: un'analisi linguistico-computazionale della variazione intralinguistica. *LIV Congresso della SLI: Corpora e studi linguistici*, 8-10 settembre 2021, Firenze: Università degli Studi di Firenze.

Mori, L., Szmrecsanyi, B. (2020). Mapping eurolects: an aggregate perspective on similarities between legislative varieties', *Languages in Contrast*, online first. https://www.jbe-platform.com/content/journals/10.1075/lic.19017.mor

Mori L. (2020a). La distribuzione dei verbi modali in testi legislativi europei e italiani. Uno studio corpus-based sulla variazione intralinguistica di dovere e potere, *AIΩN-Linguistica*, 9, 141-163.

Mori, L. (2020b). La rappresentazione di scenari deontici e l'espressione della performatività nell'italiano delle leggi: dal diritto europeo alla legislazione nazionale, *Linguistica e Filologia*, 40, 45-97.

Mori, L. (2019a) Complessità sintattica e leggibilità. Un monitoraggio linguistico per la valutazione dell'accessibilità dei testi legislativi europei e italiani. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, XLVIII (3), 627-657.

Mori, L. (2019b). Configurazioni strutturali e funzioni discorsive dei lexical bundles nella costruzione testuale delle leggi italiane nazionali e di derivazione europea. In R. Bombi, (ed.) *Saggi interlinguistici e metalinguistici*, 24, Roma, Il Calamo, 79-88.

Mori, L. (2019c) Dall'armonizzazione all'ibridazione nei testi legislativi: evidenze linguistiche e manifestazioni interculturali nell'italiano tradotti. *EntreCulturas. Revista de Traducciòn y Comunicación Intercultural*, Universidad de Málaga, 10, 377-392.

Mori, L. (2018) Observing Eurolects: The case of Italian. In L. Mori (Ed.*) Observing Eurolects. Corpus analysis of linguistic variation in EU law* (Studies in Corpus Linguistics 86), Amsterdam: John Benjamins, 200-242.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In A. Gelbukh (ed.) *International Conference on Intelligent Text Processing and Computational Linguistics*, 3-16. Springer: Cham.

Sandrelli, A. (2018) Observing Eurolects: the case of English. In L. Mori (ed.) *Observing Eurolects. Corpus analysis of linguistic variation in EU law* (Studies in Corpus Linguistics 86), Amsterdam: John Benjamins, 64-92.

Silveira, N., Dozat, T. de Marneffe, M.C., Bowman, S., Connor, M., Bauer J. & Manning C. (2014). A Gold Standard Dependency Corpus for English. In N. Calzolari *et al.* (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC-2014). 2897-2904.

Straka, M., Hajic, J., and Strakova, J (2016). Udpipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC), 4290-4297

van Halteren, H. (2004). Linguistic profiling for author recognition and verification. *Proceedings of the Association for Computational Linguistics* (ACL04), 200-207.

Venturi, G. (2013). Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach. *Bridging the Gap(s) between Language and the Law. Proceedings of the 3rd European Conference of the International Association of Forensic Linguistics*, Porto, Portugal, 15-18 October 2012, 138-156.

Vertovec, S. (2007). Super-diversity and its implications, *Ethnic and Racial Studies*, 30, 1024–1054.

Zeman, D., Nivre, J., Abrams, M. *et al.* (2019). Universal Dependencies 2.5, *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics* (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3105 .

Stylistic normalisation in translation: Differences in the use of transgressive in Czech translated and non-translated texts

Olga Nádvorníková
Charles University, Prague
olga.nadvornikova@ff.cuni.cz

Czech transgressive is a non-finite verb form conveying adverbial meanings, such as accompanying circumstance, means or manner (cf. Dvořák 1983; Nádvorníková 2010). Syntactically, it allows for condensation and hierarchisation of information in a sentence. Along with the Romance *gerundio*, the English participial -ing forms or the Russian деепричастие, Czech transgressive may be considered a converb (see Haspelmath & König 1995).

(1) a. "Bliju, soudruhu četaři," odpověděl jsem **opíraje se** rukou o zeď. (Milan Kundera, Žert (*Joke*), 1991(1969)
b. " Puking, Comrade Sergeant," I replied, **leaning** against the wall with one hand. (transl. D. Hamblyn; O. Stallybrass, 1992)
c. Je suis en train de dégueuler, camarade sergent, expliquai-je **en m'appuyant** d'une main au mur. (transl. M. Aymonin, 1975)
d. "Блюю, товарищ сержант", - ответил я, **опираясь** рукой о стену. (transl. Н. Шульгина, 1999)

In contrast with other converbs, however, the Czech transgressive is used very rarely and only in written texts, because of its archaistic stylistic mark (Cvrček et al. 2015). Archaistic character of the transgressive is the result of a normative intervention imposed to the Czech language during the national revival movement in the 19th century, taking as model of the standard literary Czech the prestigious norm of Czech texts written at the end of the 16th century. In consequence, the transgressive in standard literary Czech has a very complex morphology, although in Czech dialects and in other Slavic languages the corresponding forms went through the process of adverbialisation and their morphology was simplified. In contemporary Czech, therefore, the transgressive is no more part of the internalised, unconscious linguistic competence of speakers and its morphology has to be taught at school.

In translations into Czech, the meaning of converbs occurring in source texts is usually rendered by a (coordinate or subordinate) finite clause and the transgressive is used in only about 1-9% of cases (see Čermák et al., 2020, for translations from French, Spanish, Portuguese and Italian, and Malá & Šaldová 2015 for translations from English). A recent study (Nádvorníková 2021) conducted on the InterCorp multilingual corpus (Čermák & Rosen 2012, http://intercorp.korpus.cz) and on the Jerome comparable translation corpus of Czech (Chlumská 2013) has shown that the frequency of the transgressive is even lower in translations than in non-translated texts, which indicates the effect of stylistic normalisation (see Baker 1996, Chlumská 2017, Laviosa 2002, Lapshinova-Koltunski 2015, Toury 1995 or Vanderauwera 1985). Apart the difference between translated and non-translated texts, the aforementioned study revealed two major factors systematically influencing the frequency of the transgressive: the date of publishing of the text (the frequency of the transgressive is constantly decreasing) and the register (in fiction, the frequency of the transgressive is higher than in non-fiction).

The aim of this paper is to refine the aforementioned frequency analysis by a thorough investigation of the *use* of transgressives in order to verify the potential effect of normalisation in translations from this point of view.

Preliminary results based on the InterCorp parallel corpus (for translated texts) and on the SYNv8 reference corpus (for non-translated texts, see Křen et al. 2019) limited to fiction and non-fiction published after 1990 indicate that the use of the transgressive is more sensitive to the register than to the difference between translated and non-translated texts (see a similar observation in Kruger & van Rooy 2018 for the difference between several varieties of English): the transgressive is twice less frequent in non-fiction (both translated and non-translated) than in fiction, and specific causal meanings are prominent in non-fiction, whereas in fiction, simple temporal meanings or the prototypical accompanying circumstance dominate, especially in narrative sequences and in introductory clauses to direct speech. Moreover, half of the occurrences of transgressives in non-fiction (translated as well as non-translated) are part of citations, particularly in history books, which further reduces the frequency of the transgressive in non-fiction, observed in the previous quantitative study.

Despite these limitations, important differences between translated and non-translated texts were observed in both registers: the prototypical meaning of accompanying circumstance dominates in translated fiction as well as non-fiction, and in translated fiction the use of the transgressive is mostly limited to introductory clauses (like in (1)), in contrast with non-translated fiction, where types meanings and types of use of this form are more variable. The reduced variability of the use of the transgressive indicates the effect of normalisation in translations. This tendency may be reinforced by two facts: first, as pointed out by Nedjalkov (1995), the accompanying circumstance is the most frequent meaning conveyed by converbs in general; second, in contrast with other meanings carried by converbs, the accompanying circumstance does not have a corresponding adverbial subordinator (most European languages encode the meaning of concomitance by converbs or a simple juxtaposition of two finite clauses, see Kortmann 1997: 281).

In our presentation, we will provide more details about the corpora, the methods used for the classification of the transgressives and the statistical significance and interpretation of the differences observed. We will also investigate the potential influence the shining though effect (Teich 2003, Dai & Xiao 2011) and the impact of the target audience on the use of transgressive, i.e. the difference between translated and non-translated literature intended for children and young readers, see e.g. Čermáková 2017).

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (ed.) *Terminology, LSP and Translation: Studies in language engineering, in honour of Juan C. Sager.* Amsterdam: John Benjamins, 175–186.

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

Čermák, P., Kratochvílová, D., Nádvorníková, O. & Štichauer, P. (eds.) (2020). *Complex Words, Causatives, Verbal Periphrases and Gerund: Romance languages vs. Czech (a parallel corpus-based study)*. Praha: Karolinum. http://hdl.handle.net/20.500.11956/117388

Čermáková, A. (2017). Translating children's literature: Some insights from corpus stylistics. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies.* 71(1), 117–134.

Chlumská, L. (2013). *JEROME: srovnatelný korpus překladové a nepřekladové češtiny*. Ústav Českého národního korpusu FF UK, Praha 2013. Available at: http://www.korpus.cz

Chlumská, L. (2017). *Překladová čeština a její charakteristiky* [Translated Czech and its Characteristics]. Praha: NLN.

Cvrček, V. et al. (2015). *Mluvnice současné češtiny* [Grammar of Contemporary Czech]. Praha: Karolinum.

Dai, G. & Xiao, R. (2011). SL 'shining through' in translational language: A corpus-based study of Chinese translation of English passive. *Translation Quarterly* 62. 85–108.

Dvořák, E. (1983). *Přechodníkové konstrukce v nové češtině* [Transgressive Constructions in New Czech]. Praha: Univerzita Karlova.

Haspelmath, M. & König, E. (eds.) (1995). *Converbs in cross-linguistic perspective: Structure and meaning of adverbial verb forms – adverbial participles, gerunds*. Berlin: Mouton de Gruyter.

Kortmann, B. (1997). *Adverbial subordination: A typology and history of adverbial subordinators based on European languages*. Berlin: Mouton de Gruyter.

Křen, M. et al. (2019). *Korpus SYN, version 8 from 12/12/2019*. Praha: Institute of the Czech National Corpus. (https://www.korpus.cz)

Kruger, H. & van Rooy, B. (2018). Register variation in written contact varieties of English. *English World-Wide.* 39(2), 214–242.
Lapshinova-Koltunski, E. (2015). Variation in translation: evidence from corpora. In C. Fantionuoli & F. Zanettin (eds.) *New directions in corpus-based transaltion studies*. Berlin: Language Science Press, 93-114.
Laviosa, S. (2002). *Corpus-based Translation Studies. Theory, Findings, Applications.* Amsterdam/New York: Rodopi.
Malá, M. & Šaldová, P. (2015). English non-finite participial clauses as seen through their Czech counterparts. *Nordic Journal of English Studies* 14(1), 232–257.
Nádvorníková, O. (2010). The French gérondif and its Czech Equivalents. In *InterCorp: Exploring a multilingual corpus*. Praha: NLN. 83–95.
Nádvorníková, O. (2021). Stylistic normalisation, convergence and cross-linguistic interference in translation: The case of the Czech transgressive. In M. Bisiada (ed.) *Empirical Studies in Translation and Discourse*. Berlin: Language Science Press, 53–93. doi: 10.5281/zenodo.4450081
Nedjalkov, I. V. (1995). Some typological parameters of converbs. In M. Haspelmath & E. König (eds.) *Converbs in cross-linguistic perspective*. *Structure and meaning of adverbial verb forms – Adverbial participles, gerunds*. Berlin: Mouton de Gruyter, 97–137.
Teich, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts.* Berlin: Mouton De Gruyter.
Toury, G. (1995). *Descriptive translation studies – and beyond*. Amsterdam: John Benjamins Publishing Company.
Vanderauwera, R. (1985). *Dutch novels translated into English: The transformation of a "Minority" literature*. Amsterdam: Rodopi.

The representation of perception across languages: The French pronoun *On* and its English and Romanian translations. Evidence from a multilingual corpus[11]

Raluca Nita
Université de Poitiers
raluca.nita@univ-poitiers.fr

For many years now, the French pronoun *On* has been given special attention by linguists in the fields of both French linguistics and contrastive studies. The aim of our presentation is to confront *On*'s underspecified, vague reference (Atlani 1981, Flottum and al. 2007) with its English and Romanian equivalents as regards representation of perception in a very specific context in literary texts and show the specific pragmatic effects resulting in the three languages.

Our research is based on the GRAFE multilingual literary corpus compiled at the university of Poitiers and more specifically on the 10 excerpts from contemporary French literature (172 000 words) and their English and Romanian translations. The corpus contains 747 occurrences of *on* that have been automatically extracted with the Paraconc Concordancer and then manually described according to their French context and to their equivalents in the target languages. We have chosen to focus specifically on cases (48 in all) where *On* is the subject of a verb of visual or auditory perception and is part of the description of a scene, which is thus attached to a subjective source for which *On* stands but of which *On* doesn't allow specific identification.

> (1) La nuit descendait sur les prairies environnantes. Près de la structure de béton dans laquelle s'encastrait la chaudière, **on distinguait** une tache brunâtre imparfaitement nettoyée. (Houellebecq)
> (1') Near the concrete structure which housed the boiler, **you could make out** a brownish stain which
> had been poorly cleaned.
> (1'') Lîngă structura de beton în care era fixată centrala, **se distingea** o pată cafenie prost curăţată.

What *On* achieves here is generalization : on the one hand, what is observed or heard has a subjective origin located in the specific scene which is described, and is possibly identifiable with a specific character, and on the other hand, any other source of perception can perceive the same properties of the scene providing location within the scene. In a majority of cases, the spatial location in the scene is made explicit by adverbials of place, 23 of which are fronted (*Sur la droite… on apercevait*). It has been suggested that this syntactic specificity together with the underspecified reference of *On* and the subjective description points to *On* achieving focus on the process of perception and on the properties of the object (Hamelin 2018).

What may indeed appear as a property of this *On*-construction in a monolingual perspective can however be further developed and even questioned when adopting a contrastive approach through corpora. As will be shown, the English and Romanian equivalents either completely delete the source of perception or express generalization of perception failing to locate it relative to a specific source alongside a class of individuals as *On* does (Détrie 1998). The contrastive approach thus highlights that it is rather the location relative to a source of perception that *On* + verb of perception draws attention on.

[11] This study is part of a larger project initiated at the University of Poitiers and the Research Centre FoReLLIS on the creation of a multilingual corpus and on the analysis of *On* and its translations in English, Spanish, Romanian, German and Swedish. This study is indebted to all the members of this project Hélène Chuquet, Ramon Marti- Solano, Manuel Torrellas, Jeanne Vigneron-Bosbach, Joasha Boutault, Pauline Serpault, Ioana Daniela-Balauta, Diana Cretu, Maria Hellerstedt.

To prove our point we will discuss English and Romanian translations: the source of perception is deleted by a change in syntactic organisation where the object of perception is brought forward as a subject in active voice sentences, in passive voice sentences and *There be* constructions. When the syntactic order of the original text is maintained (subject + verb of perception), the majority of the translations have a generic pronominal subject (second person pronoun in English and Romanian, *One* in English) which could be considered to achieve a similar effect to *On*. However, different linguistic and pragmatic characteristics tend to prove otherwise: *On* can include in its fuzzy meaning reference to a third person pronoun or a first person pronoun on account of contextual interpretation (cf. Egerland 2003, Rabatel 2001), while *you* and *one* cannot. Moreover, in our translations, despite *you* and *one* filling the syntactic position of a source of perception, there is an explicit shift of focus on the properties of the scene and of the object of perception by the use of the modal *can* with the verb of perception (in 13 cases out of 14). Coupled with the adverbials of place, *can* draws the attention on the object of perception (Gilbert, 2001). Moreover, when *One* is used, context shows that the description of the scene needs to be associated to a source of perception to guarantee its properties but without this source being associated to a specific personal referent, as *On* does. The choice between *One* and *you* is attached to the characteristics of the narrative, *you* being used in novels with oral features. Thus despite the subject position of generic *one* and *you* with a perception verb, English and Romanian seem to focus on the object of perception contrary to French and its use of *On*. The corpus study thus seems to confirm an opposition between English and French in expressing perception, one focussing on the object, the other on the source of perception (Guillemin-Flescher 1994). This is all the more obvious in cases where the object of perception becomes in English subject of an active sentence or when perception is deleted from the semantics of the verb, and the location of the scene relative to a subjective origin becomes thus implicit, attached to other contextual markers.

(2) Au-dessus de la cheminée, **on voyait** un beau portrait de femme, peint à l'huile… (Simenon)
(2') Deasupra căminului **se vedea** un portret frumos de femeie, pictat în ulei…
(2'') **A fine portrait** in oils of a woman **hung** over the mantelpiece…

The choices of translation in English are very much in line with the original texts where our English to French corpus shows similar syntactic choices with similar pragmatic effects in English where French uses the *On* construction (generic *you*, passives, *there* constructions and even deletion of the perception on the whole). But back translation, however, also tends to question the choice of specific *we* in English translations, which appears to be linked to the subjectivity of the translator or even to the characteristics of the original.

In Romanian, while generic *tu* and impersonal reflexive passives are the two favoured equivalences of *On* in the whole corpus, this passive form seems to be specific to verbs of perception and to override the other two forms of verbal construction deleting the subject, prototypical passive and impersonal voice. Even though this appears to be a linguistic constraint, it contributes to conceiving the perception process as self sufficient since in this case, the grammatical subject (which is the object of perception) is in postposition (1'', 2'').

Contrastive analysis suggests specialization in relation to expressing perception in the three languages. In English and in Romanian, the focus is on the scene and the objects perceived, with a certain tendency to leave perception implicit, which seems to be difficult to achieve in French and may be more generally linked to other cases already studied like the use of *On* with opinion verbs in journalistic texts (Tartarin 2011) or even to cases of location of an object in a specific scene where *on* + *trouver* suggests the existence of a subjective observer, while English and Romanian focus again on the object located in the scene (*mais,* ***dans ce cimetière, on ne trouve*** *que cinq cent quatre-vingt-deux tombes :: but* ***the***

cemetery contains only *five hundred and eighty-two graves :: **în cimitir sînt** însă doar cinci sute optzeci şi două de morminte*.).

References

ATLANI, F. (1984). « On l'illusionniste ». In GRESILLON, A. & J-L.LEBRAVE. (Eds). La langue au ras du texte. Lille : Presses Universitaires de Lille. p. 13-29.

DÉTRIE C. (1998). « Entre ipséité et altérité : statut énonciatif de on dans Sylvie », L'information grammaticale 76, 29-33.

EGERLAND, V. (2003). "Impersonal Pronouns in Scandinavian and Romance", Working Papers in Scandinavian Syntax, 71.

FLØTTUM, K. and al. (2007). On Pronom à facettes. Bruxelles : De Boeck.

GILBERT, E. (2001). « Vers une analyse unitaire des modalités. May, must, can, will, shall», Cahiers de recherche en grammaire anglaise, T.8 Modalité et opérations énonciatives, Paris, Gap, Ophrys. 23-99

GUILLEMIN-FLESCHER, J. (1994). "Subject and Object" in YAGUELLO, M. (ed.) Subjecthood and Subjectivity, Ophrys. 171-192.

HAMELIN L. (2018) « Éléments pour une sémantique de ON ». Congrès Mondial de Linguistique Française. In SHS Web of Conferences (Vol. 46, p. 12006). EDP Sciences.

RABATEL, A. (2001). « La valeur de « on » pronom indéfini/pronom personnel dans les perceptions représentées. L'Information grammaticale, n° 88.

TARTARIN, Thérèse, 2011, Que dit ON ? ON, discours, point de vue et modalisation dans les textes journalistiques : problèmes de traduction, Mémoire de Master 2, sous la direction d'Hélène Chuquet, Université de Poitiers.

Corpus

GRAFE literary multilingual corpus, French originals to English and Romanian translations, English originals to French and Romanian translation : http://forellis.labo.univpoitiers.fr/corpus-grafe-corpus-multilingue-grec-roumain-anglais-francais-espagnol/

Translating emotions

Ulrike Oster
Universitat Jaume I
oster@uji.es

This paper is a corpus-based study of how translation affects the portrayal of emotion concepts. It tries to establish whether there are differences between translated texts and original texts in a given language as to how emotions are expressed and whether the emotion conceptualization in the translated texts is closer to that of the source language or the target language.

However different human languages, cultures and societies may be, translation is far from being an impossible task – and can even seem a rather straightforward operation – because we all share the common basis of human nature. When translators encounter cultural or linguistic differences, their strategies will move between the opposite poles of domestication (bringing the source text closer to the target reader) and foreignization (making the foreign features of the source text visible in the target text, thus bringing the reader of the translation closer to the source culture) (Venuti 1995). One particular field where the coexistence of universal and culturally constructed aspects is especially evident is that of emotions and their expression. Contrastive linguistic or anthropological studies in this field usually aim at highlighting discrepancies between supposedly equivalent emotion words (e.g. Wierzbicka 1999) or describing culture-specific, even unique emotions, which have no direct equivalents in other (especially Western) languages (e.g. Lutz 1988). In Translation Studies, we find two different approaches to the description of how translators deal with emotion-related differences. One assumes that translators will always do their best to come as closely as possible to conveying the feelings present in the source text and describes the techniques and strategies they employ (e.g. Holoborodko 2013). The other acknowledges that source and target system might be governed by differing norms and that the translator consciously or unconsciously conforms to one set of norms or the other. In this line of argument, Lamprinou (2011, 2013) studies the expression of emotion intensity in a corpus of British to Greek translations of popular romance literature and concludes that target language norms exert the stronger influence. This is contrary to the assumption of Polysystem Theory that the more powerful literary system (in this case the British) will exert the stronger influence, which means that a more foreignizing approach might have been expected.

In the cognitive linguistic tradition, which this paper adheres to, the metaphorical conceptualization of emotions has originally been understood as grounded in bodily experience and therefore universal to all human beings. However, despite many similarities, contrastive evidence has also shown that there are striking differences in the linguistic expression of emotions across languages and cultures, and also diachronic variation within the same culture (for example Gevaert 2001, 2005). One possible explanation for this is suggested by Kövecses (2005: 4). It consists in the existence of primary metaphors (such as AFFECTION IS WARMTH or CAUSES ARE FORCES), which are likely to be universal. These primary metaphors may be put together in particular languages and cultures to form “complex” metaphors, which can be language-specific. The metaphorical conceptualization of emotions is thus currently understood as being subject to the combined influence of embodiment, cognition and culture. The combination of these factors would then account for both the similarities and the differences that can be found among languages.

In order to study how the translation process affects the conceptualization of emotions, this paper focuses on one emotion in a specific language combination: the conceptual domain of ‘anger’ in German and Spanish. In the first step, an analysis of two large reference corpora (DWDS for German and Corpus del

Español for Spanish) provides a contrastive description of the concept 'anger' as represented by three prototypical emotion lexemes in both languages (Wut, Zorn, Ärger in German and ira, rabia, enojo in Spanish). The approach (Oster 2010, 2019) uses fundamental ideas from cognitive semantics, namely metaphorical and metonymical conceptualisations along with key corpus-linguistic notions like semantic preference and prosody. Through a detailed concordance analysis, it gives access to a comprehensive view of the conceptualisation of the emotion concept 'anger' in the source language (German: EmConG) and the target language (Spanish: EmConSp). It includes the following aspects: metaphorical conceptualisation; conceptual proximity (i.e. the relative position of concepts within the same conceptual domain and with respect to other concepts); physical effects of the emotion (through metonymical expressions); syntagmatic relations (which provide us with information about the prototypical causes, consequences and experiencers of the emotion); and the description and evaluation of the emotion. This phase of the study is rather advanced in the case of German through previous publications (Oster 2014) and partially so in Spanish (Oster 2019).

Once this wider picture has been established by analyzing large reference corpora, the perspective is narrowed down to the expression of 'anger' in original and translated texts in the COVALT corpus. COVALT (Valencian Corpus of Translated Literature) is a multilingual, parallel and comparable corpus containing narrative works translated from English, French and German into Spanish and Catalan as well as novels originally written in Spanish and in Catalan. For this study, the German-Spanish translation module (approx. one million words in each language) is used in combination with the original Spanish module. The following research questions guide this part of the analysis:

- Which emotion lexemes from the domain of 'anger' are present in the source texts and what translation equivalents have been chosen by the translators?
- Which aspects of the overall conceptualisation of German 'anger' (EmConG) are present in the source texts (EmConST-G)?
- What translation strategies, shifts or adaptations can be identified in the Spanish target texts?
- Is the conceptualisation of the emotion in the target texts (EmConTT-Sp) identical to that of the source texts (EmConST-G)?
- Are there changes that make EmConTT-Sp more similar to EmConSp than EmConG?

The initial analysis of the German anger expressions and their translation equivalents shows that there is considerable variability and overlap across the domain in the translation choices. Preliminary results regarding conceptual metaphors show that both source and target preferences are present in the target texts. On the other hand, a more marked deviation from target language conventions can be observed in the translation of expressions referring to physical effects or consequences of the emotion.

References

Gevaert, C. (2001). Anger in Old and Middle English: A 'Hot' Topic?. *Belgian Essays on Language and Literature* 2001, 89-101.

Gevaert, C. (2005). The ANGER IS HEAT Question: Detecting Cultural Influence on the Conceptualization of Anger through Diachronic Corpus Analysis. In N. Delbecque, J. van der Auwera & D. Geeraerts (eds.) *Perspectives on Variation: Sociolinguistic, Historical, Comparative.* Berlin & New York: Mouton de Gruyter, 195-208.

Holoborodko, A. (2013). *Some Problems of Translating Emotion Words from Russian into Japanese in F. Dostoevsky's Novel 'White Nights': Contrastive Analysis of Three Japanese Translations with the Russian Original Text Concerning Emotional Discourse*. Doctoral dissertation. Hitotsubashi University.

Kövecses, Z. (2005). *Metaphor in Culture. Universality and Variation*. Cambridge: Cambridge University Press.

Lamprinou, A. (2011). Translated Romances: The Effect of Cultural Textual Norms on the Communication of Emotions. *Journal of Popular Romance Studies 2*(1), 1-14.

Lamprinou, A. (2013). *A Study on the Cultural Variations in the Verbalisation of Near-Universal Emotions: Translating Emotions from British English into Greek in Popular Bestseller Romances*. Doctoral dissertation. University of Surrey.

Lutz, C. A. (1988). *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory*. Chicago: University of Chicago Press.

Oster, U. (2010). Using corpus methodology for semantic and pragmatic analyses: What can corpora tell us about the linguistic expression of emotions? *Cognitive Linguistics, 21*(4), 727–763.
Oster, U. (2014). Emotions between physicality and acceptability. A Contrast of the German Anger Words Wut and Zorn. *Onomázein*, 30, 286-306.
Oster, U. (2019). Cross-cultural semantic and pragmatic profiling of emotion words. Regulation and expression of anger in Spanish and German. In I. Navarro i Ferrando (ed.), *Current Approaches to Metaphor Analysis in Discourse*. Berlin & Boston: de Gruyter Mouton, 35-56.
Venuti, L. (1995). *The Translator's Invisibility. A History of Translation*. London & New York: Routledge.
Wierzbicka, A. (1999). *Emotions across Languages and Cultures*. Cambridge: Cambridge University Press.

Full reduplication as a word formation process and a translation choice. A multilingual corpus study in the context of children's literature

Cécile Poix
Université Lumière Lyon 2, CeRLA
C.Poix@univ-lyon2.fr

The aim of this presentation is to show that a translator confronted with neology can simply choose to reproduce the word-formation process of the SL, regardless of its productivity in the TL.

Neology occurs naturally in children's literature. Termed poetic occasionalisms, literary coinages can be defined as ephemeral and contextual-dependant word formations, deliberately coined for a purpose not restricted to naming. In fact, naming is rarely the main function of occasionalisms in children's literature, as they are mostly coined to entertain (Munat 2007). They can also be attention seeking devices (ASDs) (Hohenhaus 2007). They even have the power of concept formation through hypostatization (Poix 2020). From simple phonological deformation (e.g. fortin' < fortune coined by Dickens) to opaque ex-nihilo creations (e.g. mithril coined by Tolkien), the translation of occasionalisms may puzzle a translator.

Despite the alleged freedom in the translation of children's fiction (adaptation, foreignization, domestication), verbal creativity is generally reproduced, and it can simply be done by reproducing the nonce formation process. For instance, Dahl (1982) coins the occasionalism sickable < sickening by substituting the affix. Translators use the same processes in other languages: French répugnable < répugnant (Fabien: 1990) and Italian ripugnablile < ripugnante (Ziliotto: 1987). Whilst the German translator (Quidam: 1984) chose a translation by composition, etwas Magenumdrehendes [something that turns your stomach around], the substitution of an affix could have been possible: ekelsam/ekellich/ekelbar > ekelhaft. Thus, studying the nonce formation typology allows a translator to access a toolbox to neologize.

The present study focuses on one type of nonce formation process (full reduplication) which is commonly used in English for coining new words in children's literature. There are other types of reduplication, where the reduplicated element is in the initial, internal or final position. Initial reduplication is also called alliterative reduplication or reverse rhyme (e.g. squiff-squiddled coined by Dahl). Final reduplication is also called alliterative reduplication (e.g. storks-forks by Tolkien). There are two subtypes of internal reduplication: (i) the reduplicant is the repetition of one identical substring (e.g. wispy-misty coined by Dahl); (ii) the stressed vowel alternates. The latter is also known as ablaut (e.g. tip-toppling by Dalh). Full reduplication is when one word is repeated twice or more (e.g. flick-flick-flick coined by Pullman).

Generally onomatopoeic – crackety-crack, crunch-crunch, tap-tap-tapping, etc. – full reduplication provides a sound dimension to the reading of children's fictions. There are also cases of full reduplication in songs excerpts and characters' names. Occasionally, an author uses reduplication in an ex-nihilo creation (e.g. to name a fictitious entity or in an invented language).

The present study relies on a corpus entitled CHILL (CHILdren's Literature) of classics from the 19th and 20th centuries, written in and translated into English, French, German and Italian. The parallel corpus of all translations comprises over 9 million words (see CHILL in bibliography)

Using regular expressions, potential candidates were extracted from the four sub-corpora of CHILL for the analysis of full reduplication. Attested lexical units and noise were removed and the final data listed 74 full reduplicative occasionalisms.

A quantitative analysis of the reduplicative occasionalisms revealed that the nonce formation process is more common in Germanic languages (0.878) than in Romance languages (0.122). Bearing in mind that neology in literature is a type of foregrounding of the linguistic code, and that some authors are more eager than others to neologize, it is nevertheless interesting to notice the discrepancy of productivity between Germanic and Romance languages.

Translations were then observed to verify the hypothesis that reduplication is an uncommon nonce formation process for Romance languages. From the parallel corpus, five translation procedures were reviewed (Newmark: 1988): transference (keeping the SL loanword / occasionalism), naturalisation (adapting the spelling in the TL), functional equivalence, omission, and translation with a coinage using reduplication (or using reduplication already attested in the TL).

Even though full reduplication is not a productive nonce formation process in our sub-corpora of French and Italian source texts, in all four translated languages, reduplication was observed. Beside the cases of transference and adaptation, each language coined occasionalisms in translation: English (0.239), German (0.216), French (0.239), and Italian (0.306). Romance languages predominate, showing that neology in translation is not restricted to productive nonce formation processes.

Also, there are no clear patterns of formation for full reduplication. Occasionalisms seem to be arbitrarily formed of two, three or four elements, sometimes linked with stretch (reduplication of a letter). They are randomly spelt with hyphens (tick-tick), commas (tok, tok, tok, tok) typographic blanks (zac zac), stuck together (purrpurr) or combined (Schsch-Schsch). Full reduplication can also be translated by other types of reduplication (e.g. ablaut reduplication).

In the presentation, more details will be provided to show the creativity and the diversity of translation strategies as well as quantitative information.

References

Hohenhaus, P. (2007). How to do (even more) things with nonce words (other than naming). In J. Munat (ed.) *Lexical creativity, texts and* contexts. Amsterdam, Philadelphia: John Benjamins, 15-38.

Lathey, G. (2015). *Translating Children's Literature*. Routledge.

Munat, J. (2007). Lexical creativity as a marker of style in Lexical creativity, texts and contexts, In J. Munat (ed.) *Lexical creativity, texts and contexts*. Amsterdam, Philadelphia: John Benjamins,163-185.

Newmark, P. (1988). *A Textbook of Translation*. New York, London, Toronto: Prentice Hall.

Oittinen, R. (2002). Translating for children. Routledge.

Paloposki, O. (2011). Domestication and foreignization. *Handbook of translation studies*, 2, 40-42.

Poix, C. (2020). L'hypostatisation des occasionnalismes poétiques dans la littérature pour la jeunesse, ou l'innovation lexicale suffit-elle à créer un concept ? *Neologica* N°14., 145-166

Schmid, H. J. (2016) [2005]. *English morphology and word-formation* (3rd ed.). Erich Schmidt Verlag.

Van Coillie, J., & Verschueren, W. P. (2014). *Children's literature in translation: Challenges and strategies*. Routledge.

Vinay, J. P., & Darbelnet, J. (1995). *Comparative stylistics of French and English: A methodology for translation* (Vol. 11). John Benjamins Publishing.

CHILL: Parallel corpus of CHILdren's Literature:

- English:

Barrie, J.M. (1911). *Peter Pan (Peter and Wendy)*. Translations DE Skunca (n.d.), FR Laporte (2009), IT Dandolo (n.d.).

Carroll, L. (1865). *Alice's adventures in Wonderland*.Translations DE Zimmermann (1869), FR Bué (1869), IT Pietrocòla-Rossetti (1872).

Dahl, R. (1982). *The BFG*. Translations DE Quidam (1984), FR Fabien (1990), IT Ziliotto (1987).

Dickens, C. (1837). *Oliver Twist*. Translations DE Kolb & Ritthaler (1971), FR Gérardin & Lorain (1893), IT Amato (2014).

Kipling, R. (1894). *The Jungle Book*. Translations DE unknown (n.d.), FR Fabulet & d'Humières (1899), IT Dauli (2003?).

Lewis, C.S. (1950). *The Lion, the Witch and the Wardrobe*. Translations DE Tetzner (1981), FR Dalmais (2005), IT Dei (1979).

Pullman, P. (1995). *Northern Lights*. Translations DE Ströle & Kann (2002), FR Esch (1998), IT unknown (n.d.).

Rowling, J.K. (1997). *Harry Potter and the Philosopher's Stone*. Translations DE Fritz (1998), FR Ménard (1998), IT Astrologo (1998).
Stevenson, R. L. (1882). *Treasure Island*. Translations DE unknown (n.d.), FR Serval (1920), IT unknown (n.d.).
Tolkien, J.R.R. (1937). *The Hobbit*. Translations DE Krege (1998), FR Ledoux (1980), IT Società Tolkieniana Italiana (n.d.).
Wilde, O. (1887). *The Canterville Ghost*. Translations DE unknown (1993), FR Robillot & Couvin (199), IT unknown (n.d.).

- French:

Alain-Fournier (1913). *Le grand Meaulnes*. Translations DE Langrebe (2009), EN Russel (1999), IT Bianconi (1998).
Aymé, M. (1934). *Les contes du chat perché*. Translations DE Meier-Haas & Lang (1964), EN Denny (1951), IT Galeotti & Lazzaro (2005).
Daudet, A. (1869). *Lettres de mon moulin*. Translations DE Kühne (1900), EN Harmelin & Adams (2009), Dandolo (1945).
Dumas, A. (1845). *Le Comte de Monte-Cristo*. Translations DE Pannwitz, Rütten & Loening (n.d.), EN unknown (1888), IT unknown (n.d.).
Giono, J. (1953). *L'homme qui plantait des arbres*. Translations DE Tappolet (n.d.), EN Doyle (n.d.), IT Spagnol (n.d.).
Hugo, V. (1831). *Notre-Dame de Paris*. Translations DE unknown (n.d.), EN Hapgood (n.d.), IT unknown (n.d.).
Ponti, C. (1998). *Ma vallée*. Translations DE Klewer (1999), EN Waters (2017), IT Gandini (2001).
Prévert, J. (1945). *Paroles*. Translations DE Kusenbergs (1962), EN Ferlinghetti (1958), IT Cortiana (n.d.).
Saint-Exupéry, A. (1943). *Le petit prince*. Translations DE Leitgeb (1956), EN Howard (2005), IT unknown (n.d.).
Sand, G. (1848). *La petite Fadette*. Translations DE Moellenhoff (2007), EN Mrc (2013), IT Calvanese (2014).
Verne, J. (1869). *Vingt mille lieues sous les mers*. Translations DE Jürgensmeister (n.d.), EN Walter (2001), IT unknown (n.d.).

- German:

Eichendorff, J. (1826). *Aus dem Leben eines Taugenichts*. Translations EN Glenny (n.d.), FR Laval & Sctrick (1990), IT unknown (n.d.).
Ende, M. (1973). *Momo*. Translations EN Brownjohn (1985), FR Gepner (2009), unknown (n.d.).
Grimm, J. & W. (1812). *Kinder- und Hausmärchen*. Translations EN Hunt (1884), FR unknown (n.d.), IT Gramsci (1980).
Hoffmann, E.T.A. (1816). *Nussknacker und Mausekönig*. Translations EN Neugroschel (2007), FR de la Bédollière (1838), IT Peltenburg-Brechneff (2011).
Hoffmann, H. (1845). *Der Struwwelpeter*. Translations EN unknown (n.d.), FR Trim (1872), IT Negri (1882).
Kästner, E. (1928). *Emil und die Detektive*. Translations EN Halll (1959), FR Georges (2007), IT Mazzucchetti (1931).
Nöstlinger, C. (1975). *Konrad oder das Kind aus der Konservenbüchse*. Translations EN Bell (1986), FR Royer (1982), Calamai (2007).
Preussler, O. (1967). *Das kleine Gespenst*. Translations EN Bell (2001), Kahn (1979), IT Fischer (2007).
Salten, F. (1923). *Bambi: eine Lebensgeschichte aus dem Walde*. Translations EN unknown (n.d.), FR Waquet (2016), Pecchi (2015).
Spyri, J. (1881). *Heidi*. Translations EN Pausinger (2007), FR unknown (1882), IT Lamberti (2016).

- Italian:

Amicis, E. (1886). *Cuore*. Translations DE Freese (1986), EN Hapgood (n.d.), FR Piazzi (n.d.).
Buzzati, D. (1945). *La famosa invasione degli orsi in Sicilia*. Translations DE Ringe (2000), EN Lobb (2016), FR Pasquier (1968).
Calvino, I. (1963). *Marcovaldo; ovvero, Le stagioni in città*. Translations DE Riedt & Erné (1988), EN Weaver (1983), FR Stragliati (1979).
Collodi, C. (1883). *Le avventure di Pinocchio*. Translations DE Grumman (n.d.), EN Carsey (n.d.), FR Sartirano (1883).
Piumini, R. (1993). *Mattia e il nonno*. Translations DE Bucholz (2011), EN Quigly (1993), FR Ménard (1994).
Rodari, G. (1978). *C'era due volte il Barone Lamberto*. Translations DE Scholl (1983), EN Shugaar (2011), FR Salomon (2007).
Salgari, E. (1900) [1883-1884]. *Le tigri di Mompracem*. Translations DE Wurm (n.d.), EN Lorenzutti (2014), FR unknown (1900).

A Comparison of Discourse Particles in English Original and Simultaneous Interpreted Speeches

Christina Pollkläsener
Universität des Saarlandes
s8pochri@stud.uni-saarland.de

Even though Discourse Particles (henceforth DPs) are very typical for spoken language (cf. e.g. Ajimer 2002:2), they are an understudied phenomena in interpreting. The only studies to my knowledge are Defrancq (2016, 2018) and Bendazzoli (2019) who found differences in the usage of the DPs "well", "now" and "so" between simultaneous interpretations and original speeches. Building on these findings, the research question of this study was: What are similarities and differences in the usage of the Discourse Particles "well" and "now" in spoken original and simultaneous interpreted speeches in English in the European Parliament? The DPs "well" and "now" were chosen for this purpose as they share some distributional properties and preliminary analysis showed their relatively high incidence in the corpus.

To answer this question, a corpus-based approach was taken and quantitative as well as qualitative analyses were carried out. The EPIC UdS Corpus was used which contains English simultaneously interpreted and original MEP speeches from the European Parliament. The source language of the interpretations was German. The corpus was queried for "well" and "now" and DP uses were separated from adverbial ones. To do this, the online query tool CQPweb and its "categorise" function were employed. This function allows to create categories and to manually classify the query results into these user-defined categories. In additional steps, DP functions were manually annotated and interpretations were manually aligned with the corresponding segments in the German source.

The quantitative analysis showed that there was no significant difference ($p>0.05$) between the occurrence of "well" in interpreting (52 per 100,000 token) and original speeches (62 per 100,000 token). However, "now" was more than twice as frequent in interpreting (164 per 100,000 token) than in originals (74 per 100,000 token), which was significant ($p<0.05$). Only a fraction of the DPs in the interpreting sub-corpus (8% for "well", 5% for "now") corresponded to a DP in the German source and were considered triggered. This was even lower than in Defrancq (2016: 115, 2018: 126), substantiating the notion that interpreters use DPs independently.

For the qualitative analysis, a new framework of DP functions that sought to capture differences between interpreters' and MEP's usage was developed. Based on Ajimer (2002, 2011) and considering that the register of political speeches is conducive to the use of markers that structure argumentative steps as well as to markers that signal value judgements, the framework distinguishes between frame (textual, e.g. direct speech introduction, topic-change) and attitude (e.g. disagreement, shift to evaluation) macrofunctions. As interpreting is a highly stressful activity (AIIIC, 1999, as cited in Blumenthal et al., 2006: 483) and because Crible (2018) identified three DP functions that directly signal disfluent speech (e.g. self-repair, stalling), a stress macrofunction was introduced, creating a novel framework that is uniquely capable to measure the influences on interpretation.

Table 3: Distribution of macrofunctions for "now" and "well" in the two sub-corpora

	Original *well*		**Interpreted *well***		**Original *now***		**Interpreted *now***	
	Abs.	*%*	*Abs.*	*%*	*Abs.*	*%*	*Abs.*	*%*
Frame	19	51	24	62	42	79	80	78
Attitude	17	46	6	15	10	19	19	19
Stress	1	3	9	23	1	2	3	3
Total	**37**	**100**	**39**	**100**	**53**	**100**	**102**	**100**

The qualitative analysis involved two steps. First, the distribution of the macrofunctions for “well” and “now” between original and interpreted speeches were compared. Then, a fine-grained analysis was carried out, where the textual context was taken into account and incidences of individual functions within the macrofunctions were compared. The distribution among macrofunctions was significantly different ($p<0.05$) between “well” in interpretations and originals. MEP employed “well” almost equally often for frame macrofunctions (51%) as for attitude ones (46%), leaving stress macrofunctions by far as the least frequent group (3%). Interpreters also often used “well” for frame uses (62%), but showed a lower inclination to employ it for attitude functions (15%) and a higher inclination towards stress macrofunctions (23%). There was no significant difference ($p>0.05$) between original and interpreted speeches regarding distribution among the macrofunctions for “now”. In both sub-corpora, the frame macrofunction was by far the most frequent (around 79% for both), followed by attitude (19% for both) and stress was the least frequent one (around 2% for both).

The fine-grained, qualitative analysis gave context to the analyses presented above. It was found that MEP and interpreters used similar structures to realise conversational routines involving “well” and “now” like initiation, direct speech introduction, prefacing the answer to a question and transition. In these cases, the interpreter adequately rendered into English the German sequence that corresponded to one of these conversational routines. Differences in the frequency of the realisation of these structures hinted at stylistic differences between European Parliament speeches in English and German. For example, the frequency for “well” introducing direct speech in the interpreting sub-corpus (18%) was higher compared to original speeches (8%) as well as higher compared to what Defrancq (2016) found in his interpreting sub-corpus (9%), where the source languages were Spanish, French and Italian. This suggests that German-speaking MEP make more use of direct speech as a stylistic device than English-speaking ones, which could be a cross-linguistic register difference.

Some uses of “well” and “now” were found to be peculiar to interpreting. These uses usually involved contexts of stress and high cognitive load. On the one hand, there was a category in the stress macrofunction called “semantic gap” that captured instances, where the interpreter is struggling and inserts a DP to cover up an omission. Defrancq (2016) already found this use for “well”. On the other hand, stress also coloured a lot of uses within the frame macrofunction for “well” and “now”. Whereas the cases where MEP used “now” for topic-change were quite clear, interpreters’ use of topic-change “now” often had a flimsy quality. Considering that “now” had a significant ($p<0.05$) higher frequency in interpreted speeches compared to originals, flimsy topic-change “now” could be seen as a characteristic attribute of interpreted speeches. Frame “well” was found as part of padding and chunking techniques that interpreters use to deal with cognitive load, either as a way to cover up omissions (cf. Gumul, 2017) or to reduce syntactic complexity (cf. Seeber, 2011).

References

Ajimer, K. (2002). *English Discourse Particles: Evidence From a Corpus.* Amsterdam: John Benjamins Publishing Company. DOI: https://doi.org/10.1075/scl.10

Ajimer, K. (2011). Well I'm not sure I think… The use of well by non-native speakers. *International Journal of Corpus Linguistics,* 16(2), 231-254. DOI: 10.1075/ijcl.16.2.04aij

Aijmer, K., (2013). *Understanding Pragmatic Markers: A Variational Pragmatic Approach*. Edinburgh: Edinburgh University Press.

Bendazzoli, C. (2019). Discourse markers in English as a target language: the use of so by simultaneous interpreters. *Textus*, 32(1):183-201.

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. New Frontiers in Translation Studies* (pp. 21-42). Singapore: Springer, Singapore. DOI: https://doi.org/10.1007/978-981-10-6199-8_2

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

Blakemore, D. (2002). *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511486456

Blumenthal, P., Britt, T. W., Cohen, J.A., McCubbin, J., Maxfield, N., Michael, E. B., Moore, P., Obler, L.K., Scheck, P., Signorelli, T.M. & Wallsten, T.S. (2006). Stress effects on bilingual language professionals' performance. *International Journal of Bilingualism*, 10:(4), 477-497. DOI: 10.1177%2F13670069060100040501

Blühdorn, H., Foolen, A. & Loureda, O. (2017). Diskursmarker: Begriffsgeschichte – Theorie – Beschreibung Ein bibliographischer Überblick. In H. Blühdorn, A. Deppermann, H. Helmer & T. Spranz-Fogasy (Eds.) Diskursmarker im Deutschen. Reflexionen und Analysen (pp. 7-48). Göttingen: Verlag für Gesprächsforschung. Retrieved from: http://www.verlag-gespraechsforschung.de/2017/pdf/diskursmarker.pdf

Crible, L. (2018). *Discourse Markers and (Dis)fluency: Forms and Functions Across Languages and Registers*. Amsterdam: John Benjamins Publishing Company. DOI: https://doi.org/10.1075/pbns.286

Defrancq, B. (2016). Well, interpreters… a corpus-based study of a pragmatic particle used by simultaneous interpreters. In G. Corpas Pastor & M. Seghiri Dominguez (Eds.), *Corpus-based approaches to translation and interpreting: from theory to applications* (Vol. 106, pp. 105–128). Bern: Peter Lang.

Defrancq, B. (2018). The European Parliament as a discourse community: its role in comparable analyses of data drawn from parallel interpreting corpora. *The Interpreters' Newsletter*, 23, 115–132. DOI: 10.13137/2421-714X/22401

Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics*, 14(3), 383-398. DOI: https://doi.org/10.1016/0378-2166(90)90096-V

Fraser, B (1999). What are discourse markers? *Journal of Pragmatics*, 31(7), 931-952. DOI: 10.1016/S0378-2166(98)00101-5

González, M. (2005). Pragmatic Markers and Discourse Coherence Relations in English and Catalan Oral Narrative. *Discourse Studies*, 77:(1), 53–86. DOI: 10.1177/1461445605048767

Gumul, E. (2006). Explicitation in Simultaneous Interpreting: a Strategy or a By-Product of Language Mediation? *Across Languages and Cultures*, 7:(2),171–190. DOI: 10.1556/Acr.7.2006.2.2

Gumul, E. (2017). Explicitation and Directionality in Simultaneous Interpreting. *Linguistica Silensiana*, 38, 311-329. Retrieved from: https://www.researchgate.net/publication/322255476

Hooper, P. (1991). On some Principles of Grammaticalization. In E.C. Traugott & B. Heine (Eds). *Approaches to Grammaticalization: Volume I Focus on Theoretical and Methodologocial Issues* (pp. 17-35). Amsterdam/Philadelphia: John Benjamins Publishing Company. Retrieved from: https://www.researchgate.net/publication/284156121_On_some_Principles_of_Grammaticalization

Imo, W. (2017). Diskursmarker im gesprochenen und geschriebenen Deutsch. In H. Blühdorn, A. Deppermann, H. Helmer & T. Spranz-Fogasy (Eds.), *Diskursmarker im Deutschen. Reflexionen und Analysen* (pp. 49-72). Göttingen: Verlag für Gesprächsforschung. Retrieved from: http://www.verlag-gespraechsforschung.de/2017/pdf/diskursmarker.pdf

Klaudy, K. (2011). Explicitation. In M. Baker & G. Saldanha (Eds.), *Routledge Encyclopedia of Translation Studies* (2nd ed.). New York: Routledge.

Magnifico, C. & Defrancq, B. (2020). Norms and gender in simultaneous interpreting: a study of connective markers. *The International Journal of Translation and Interpreting Research*, 12:(1), 1-17. DOI : 10.12807/ti.112201.2020.a01

Redeker, G. (1991). Linguistic Markers of Discourse Structure. *Linguistics*, 29, 1139–1172. DOI: 10.1515/ling.1991.29.6.1139

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511611841

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories - New models. *Interpreting*, 13:(2), 176–204. DOI: 10.1075/intp.13.2.02see

Creating a new-generation corpus for corpus-based translation studies: the case of Dutch Parallel Corpus 2.0

Ryan Reynaert, Gert De Sutter
Ghent University
ryan.reynaert@ugent.be, gert.desutter@ugent.be

Since the early 1990s, the availability of translational corpora has extensively contributed to the identification of the specific linguistic features of translated text in comparison to their source texts and comparable non-translated texts (see De Sutter & Lefer 2019 for a critical overview). Nevertheless, recent empirical translation studies (e.g., Halverson 2015, De Sutter et al. 2012, Lefer 2020) agree that most traditional corpora are no longer suited to account for a thorough understanding of the (cognitive and social) mechanisms that shape the language used in translated texts. In order to further uncover the sociocognitive circumstances under which texts and translation are produced, compilers of (parallel) corpora are encouraged to develop new-generation corpora which are "more carefully designed to take consideration of translators' backgrounds and the circumstances of text production" (Kotze 2020: 356).

The present research project responds to this invitation for more qualitative data in corpus-based translation studies by introducing the Dutch Parallel Corpus 2.0 (DPC 2.0): a bidirectional parallel corpus of expert translations for Dutch><English and Dutch><French language pairs. The corpus, which readopts the main compilation and design principles of its predecessor (Macken, De Clercq & Paulussen 2011), at the time of writing contains 2.75 million words and is furthermore sentence-aligned, lemmatized and POS-tagged by means of the state-of-the-art natural language processing toolkit Stanza. DPC 2.0 distinguishes itself from traditional parallel corpora through its considerable amount of metadata about the translators (e.g., gender, education, experience) and the translation projects (e.g., L1/L2 translation, software used, degree and type of revision), next to the traditional metadata about the texts and translations themselves (e.g., source and target language, intended audience, intended goal, register).

One of the main challenges in this corpus compilation project was to adopt a principled yet flexible approach to register classification. This entailed a bottom-up approach, in which all texts in the corpus were annotated for four situational characteristics – *text provider, intended audience, channel/mode* and *communicative purpose* (Biber 1994, Delaere 2015 and Delaere & De Sutter 2017) – which serve as the basis for the eventual register classification. This annotation process was performed independently by two main annotators as well as multiple student annotators who were hired in the case of hybrid texts containing for instance more than one communicative purpose. This means that for each text in the corpus, an interannotator agreement score is available for each of the situational characteristics. Depending on the specific research goals, each user can create his/her own register classification based on one or more of the situational characteristics or decide to leave out certain texts with a low agreement score on crucial characteristics. In creating such a personalized register classification, researchers should obviously always be aware that the selection of only one or two situational characteristics automatically leads to registers at a low level of specificity and are expected to ascertain a wide range of linguistic variety within their registers, and vice versa. Corpus users can also decide to adopt the register classification by the corpus compilers themselves, who used each of the four situational characteristics to establish the nine registers: *manuals for a general audience, manuals for specialists, (popular) science, journalistic texts, commercial communication, public service communication, political speeches, literature and touristic texts*.

In advocating a bottom-up approach to register classification, we deal with two prevailing issues concerning traditional top-down register labels. First, many corpora are still essentially organized in terms of a shared topic or a shared domain of use (e.g. *history, science, legal*). Whereas topic-related typologies

conveniently point towards differences in, among others, vocabulary choice, they do not fully cover potential linguistic and situational heterogeneity within a single text category nor do they consider potential homogeneity across categories (Biber & Conrad, 2009). For instance, in the first version of the DPC (Macken, De Clercq & Paulussen 2011), topic-related texts such as *scientific debates* and *scientific articles* were clustered together, despite their varying production circumstances (e.g. spoken and written mode) and expected linguistic differences. On the other hand, predictable linguistic similarities between *scientific articles* and *historical articles* remained unnoticed. Second, although corpora such as the version of the DPC make an initial attempt to cluster texts according to six so-called 'text types', a.o. *external communication, fictional literature* and *instructive texts*, these top-down labels usually give rise to a cluster of heterogeneous texts which are defined at different levels of specificity. In fact, whereas *external communication* clusters texts according to their shared intended (broad) audience, *instructive texts* are more concretely defined in function of a shared communicative purpose, regardless of the intended audience.

DPC 2.0 thus allows researchers from various disciplines to adopt a fine-grained approach to linguistic research on translations and their source texts in which the underlying, extra-linguistic context plays an important role. The output of each search query can in fact be filtered according to a large variety of text-related, translation-related and translator-related criteria, as well as a flexible combination of multiple criteria. As a result, end-users of DPC 2.0 are enabled to carry out descriptive-comparative analyses of, for instance, varying translator profiles or translational contexts.

References

Biber, D. (1994). An analytical framework for register studies. In D. Biber & E. Finegan (ed.), *Sociolinguistic Perspectives on Register* (pp. 31--56). Oxford University Press.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

Delaere, I. (2015). *Do translations walk the line?: visually exploring translated and non-translated texts in search of norm conformity*. Ghent University. Faculty of Arts and Philosophy, Ghent, Belgium.

Delaere, I., & De Sutter, G. (2017). Variability of English loanword use in Belgian Dutch translations : measuring the effect of source language, register, and editorial intervention. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies : new methodological and theoretical traditions* (Vol. 300, pp. 81–112). Berlin / Boston: De Gruyter Mouton.

De Sutter, G., Goethals, P., Leuschner, T., & Vandepitte, S. (2012). Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures, 13*(2), 137–143.

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-studies in translation theory and practice, 28*(1), 1–23.

Halverson, S. L. (2015). Cognitive Translation Studies and the Merging of Empirical Paradigms. The Case of 'literal Translation.' *Translation Spaces*, *4*(2), 310–40.

Kotze, H. (2020). Converging what and how to find out why: An outlook on empirical translation studies. In L. Vandevoorde, J. Daems & B. Defranq (Eds.), *New Empirical Perspectives on Translation and Interpreting* (pp. 333-371). Routledge.

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

Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch parallel corpus: a balanced copyright-cleared parallel corpus. *META*, *56*(2), 374–390. https://doi.org/10.7202/1006182

Analysing the Phraseology of Named Bays for their Representation in a Terminological Knowledge Base

Juan Rojas-Garcia
University of Granada
juanrojas@ugr.es

EcoLexicon (http://ecolexicon.ugr.es) is an electronic, multilingual, terminological knowledge base on environmental sciences that is the practical application of Frame-based Terminology (Faber, 2012). Its flexible design permits the contextualization of data so that they are more relevant to specific subdomains and geographic areas (León-Araúz et al., 2013). However, to ease the geographic contextualization of concepts such as those belonging to the semantic category of LANDFORM, it is necessary to know which terms are semantically related to each named landform (e.g., *Salinas River*, *Chesapeake Bay*, *Sunset Beach*), and how those terms are linked to each other.

Although named landforms, among other named entities, are frequently found in specialized texts on environment, their representation and inclusion in knowledge resources have received little research attention. So far, knowledge resources have limited themselves to representing concepts such as BAY, RIVER or BEACH, on the questionable assumption that the concepts linked to each of them are also related, respectively, to all named bays, rivers and beaches in the real world. This issue is evident in the following description of forcing mechanisms acting on suspended sediment concentrations (SSC) in bays and rivers.

According to Moskalski & Torres (2012), temporal variations in the SSC of bays and rivers are the result of a variety of forcing mechanisms. River discharge is a primary controlling factor, as well as tides, meteorological forcing (i.e., wind-wave resuspension, offshore winds, storm, and precipitation), and human activities. Several of these mechanisms tend to act simultaneously. Nonetheless, the specific mix of active mechanisms is different in each bay and river. For example, SSC in *San Francisco Bay* is controlled by spring-neap tidal variability, winds, freshwater runoff, and longitudinal salinity differences, whereas precipitation and river discharge are the mechanisms in *Suisun Bay*. In *Yangtze River*, SSC is controlled by tides and wind forcing, whereas river discharge, tides, circulation, and stratification are the active forcing mechanisms in *York River*.

Consequently, in a knowledge resource, a list of forcing mechanism concepts semantically linked to BAY and RIVER concepts would not accurately represent the knowledge really transmitted in specialized texts. To cope with this type of situation, terminological knowledge bases should include the semantic description of named landforms by analysing the specific phraseology that surrounds them in specialized texts. In addition, recent research has shown that the linguistic behaviour of proper names may differ from that of common nouns (Helmbrecht et al., 2018; Nübling et al., 2015; Schlücker & Ackermann, 2017).

This paper thus analyses the specialized phraseology used in relation to colponyms (i.e., named bays such as *Monterey Bay*) in Coastal Engineering texts, and describes the semantic annotation of the predicate-argument structure of sentences where a colponym is mentioned. The sentences were extracted from a subcorpus of English texts on Coastal Engineering, comprising roughly 7 million tokens and composed of specialized texts (scientific articles, technical reports, and PhD dissertations), and semi-specialized texts (textbooks and encyclopaedias on Coastal Engineering). This subcorpus is part of the English EcoLexicon Corpus (23.1 million tokens) (see León-Araúz et al. (2018) for a detailed description). The automatic detection of the colponyms in the corpus was performed with a GeoNames database dump (http://www.geonames.org). GeoNames has over 10 million proper names for 645 different geographic

entities, such as bays, beaches, rivers, and mountains. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored.

A set of 1,603 sentences, corresponding to 294 different bays (see Figure 1), were annotated by two terminologists in the INCEpTION annotation tool (Klie et al., 2018) (see Figure 2). The inter-annotator agreement measure was *Cohen's kappa* (henceforth referred to as κ). The elements annotated were the following:

(1) The semantic category of the arguments, based on those implemented in EcoLexicon (Gil-Berrozpe et al., 2019), namely a hierarchically-organized list of 152 semantic categories distributed in up to five categorization levels (e.g., the semantic category of the BEACH-SIZE SAND concept is MINERAL, placed on the fifth level of the category hierarchy ENTITY>MATTER>SOLID MATTER>MATERIAL>MINERAL). The inter-annotation agreement was κ=93%, and *p*-value<0.05.

(2) The semantic role of the arguments (e.g., AGENT, PATIENT, THEME, *inter alia*). The inter-annotation agreement was κ=91%, and *p*-value<0.05.

(3) The semantic relation held between the arguments, based on those in EcoLexicon (Faber et al., 2009) (e.g., *takes_place_in*, *located_at*, *attribute_of*, *causes*, and *affects* among others), with the addition of specific relations in the context of colponyms (e.g., *deposits*, *drains*, *moves_over*, among others). The inter-annotation agreement was κ=92%, and *p*-value<0.05.

(4) The lexical domain of the verbs (and their nominalizations), based on the classification into eight domains proposed by Faber & Mairal (1999) within the framework of the Lexical Grammar Model, namely EXISTENCE (e.g., *be*, *happen*), POSITION (e.g., *put*), CHANGE (e.g., *become*, *change*), POSSESSION (e.g., *have*), MOVEMENT (e.g., *go*, *move*), MANIPULATION (e.g., *use*), ACTION (e.g., *make*), and COGNITION (e.g., *know*). The inter-annotation agreement was κ=82%, and *p*-value<0.05.

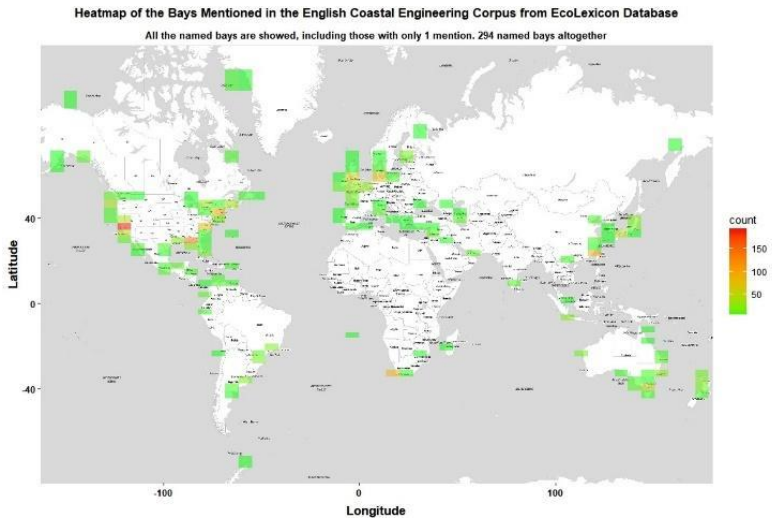


Figure 1: Location of the named bays in the English Coastal Engineering Corpus from Ecolexicon Database.

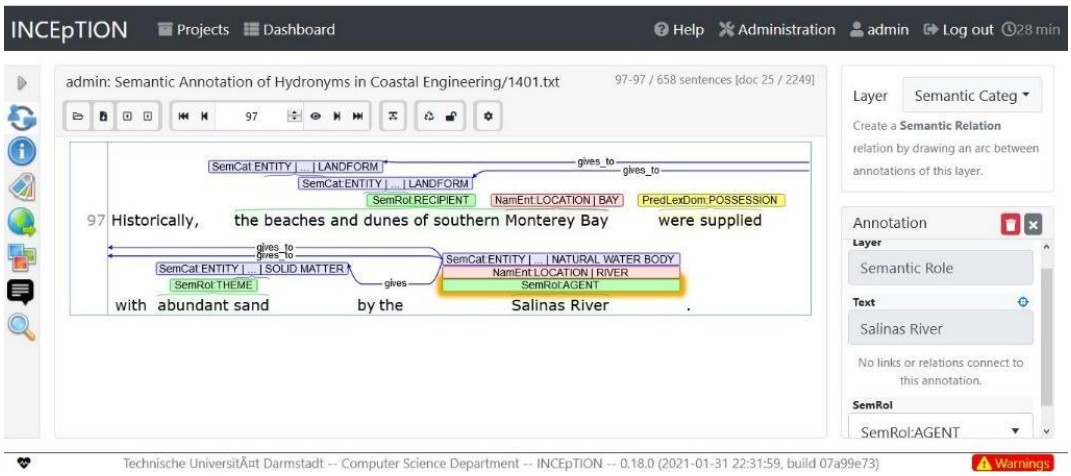


Figure 2: Annotation of the sentences where hydronyms, such as named bays and rivers, are mentioned in the Coastal Engineering corpus with the INCEpTION annotation tool.

The results, on the one hand, allowed us to draw conclusions on how each lexical domain of the verbs (and their nominalizations) employed in the context of colponyms was configured, namely the specific combination of semantic roles and categories, and the semantic relation conveyed by their different patterns of combination. The percentages of annotated sentences classified into the predicate lexical domains are shown in Figure 3, which reflects that the lexical domains of POSSESSION, MOVEMENT, and CHANGE encompassed almost 60% of the sentences.

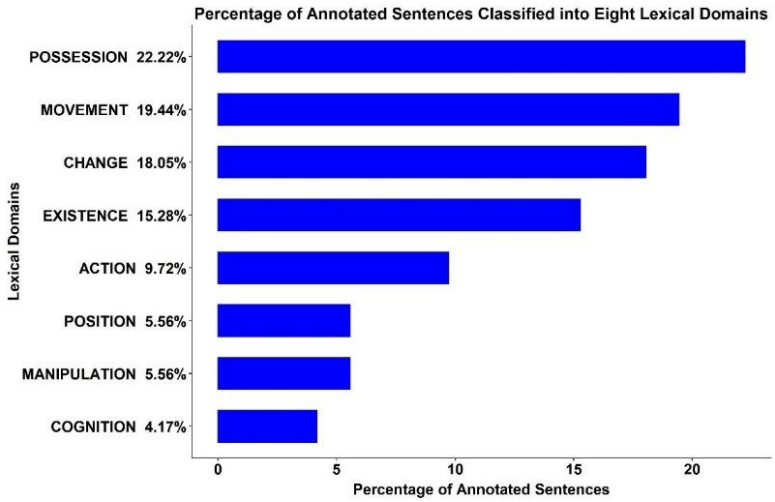


Figure 3: Percentage of annotated sentences classified into lexical domains.

For instance, Figure 4 summarizes the findings for the lexical domain of MOVEMENT (19.44% of the sentences). Only two different combinations of semantic roles were found: (1) AGENT + PATIENT, and (2) THEME + LOCATION. The second pattern (THEME + LOCATION) always conveyed the *moves_into* semantic relation, whereas the first one (AGENT + PATIENT) could express two relations, namely *drains,* or *moves_over*. However, the semantic relation transmitted by the first pattern could be always discriminated, thanks to the semantic category of the concept playing the PATIENT role. As such, the first pattern encoded the *drains* relation if the concept with the PATIENT role belonged to the LANDFORM category (e.g., *watershed*), while the pattern conveyed the *moves_over* relation if the concept belonged to the PART OF WATER BODY category (e.g., *bank*).

Verb Lexical Domain: MOVEMENT						
Arg 1	Arg 2	Arg 3	Example	Term	Relation	Term
AGENT Named Bay	PATIENT ENTITY > LANDFORM		San Francisco Bay drains large watersheds.	San Francisco Bay	drains	watershed
Named Bay	ENTITY > PART > PART OF WATER BODY		The Monterey Bay overflows its banks and deposits sediments in the flood plain.	Monterey Bay	moves_over	bank
THEME ENTITY > MATTER > SOLID MATTER	LOCATION Named Bay		Thus, several tens of millions of bed load goes into the Pensacola Bay along with the ebb currents.	bed load	moves_into	Pensacola Bay

Figure 4: Summary of the findings for the verb lexical domain of MOVEMENT, namely the combination of semantic roles and categories of the arguments in the analysed sentences, and the three semantic relations held between colponyms, and other terms mentioned in the sentences.

On the other hand, since language is a conceptual mirror that reflects how specialized knowledge is structured (Faber & Cabezas-García, 2019), the analysis of the phraseology permitted us to represent specialized knowledge of colponyms in semantic networks in EcoLexicon, according to the theoretical premises of Frame-based Terminology. For instance, the semantic frame in Figure 5 underlies the linguistic usage of *Monterey Bay* (in California, the USA) in Coastal Engineering texts, and makes the semantic and syntactic behavior of terms explicit by means of the description of conceptual relations and term combinations (Faber, 2009).

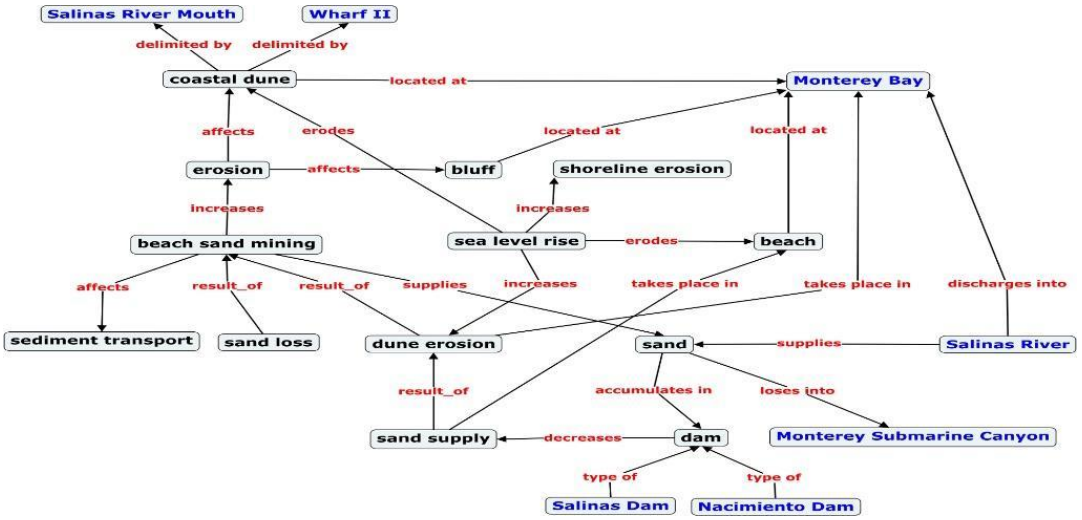


Figure 5: Semantic networks of the terms associated with *Monterey Bay* (California) in Coastal Engineering texts.

Finally, the statistical analysis of the annotations, applying the machine learning techniques of *association rules* and *decision trees*, revealed which rules permit the prediction of certain semantic annotations, a fact that could be beneficial to the implementation of automatic semantic annotators.

References

Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. *MonTI. Monografías de Traducción e Interpretación* 1, 107-134.

Faber, P. (ed.). (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin & Boston: De Gruyter Mouton.

Faber, P. & Cabezas-García, M. (2019). Specialized Knowledge Representation: From Terms to Frames. *Research in Language* 17(2), 197-211.

Faber, P. & Mairal, R. (1999). *Constructing a Lexicon of English Verbs*. Berlin & New York: Mouton de Gruyter.

Faber, P., León-Araúz, P., & Prieto, J.A. (2009). Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies* 1, 1-23.

Gil-Berrozpe J.C, León-Araúz, P., & Faber, P. (2019). Ontological Knowledge Enhancement in EcoLexicon. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.)

Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference. Sintra: Lexical Computing CZ, 177-197.

Helmbrecht, J., Denk, L., Thanner, S., & Tonetti, I. (2018). Morphosyntactic Coding of Proper Names and its Implications for the Animacy Hierarchy. In S. Cristofaro & F. Zúñiga (eds.) *Typological Hierarchies in Synchrony and Diachrony*. Amsterdam: John Benjamins, 381-404.

Klie, J.C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Santa Fe (New Mexico, USA): ACL, 5-9.

León-Araúz, P., Reimerink, A., & Faber, P. (2013). Multidimensional and Multimodal Information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem & P. Fuglewicz (eds.) *Computational Linguistics*. Berlin: Springer, 143-161.

León-Araúz, P., San Martín, A., & Reimerink, A. (2018). The EcoLexicon English Corpus as an open corpus in Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress*. Ljubljana: Euralex, 893-901.

Moskalski, S. & Torres, R. (2012). Influences of Tides, Weather, and Discharge on Suspended Sediment Concentration. *Continental Shelf Research* 37, 36-45.

Nübling, D., Fahlbusch, F., & Heuser, R. (2015). *Namen: Eine Einführung in die Onomastik*. Tübingen: Narr.

Schlücker, B. & Ackermann, T. (2017). The Morphosyntax of Proper Names: An Overview. *Folia Linguistica* 51(2), 309-339.

Gender bias and Machine Translation: *On first looking into parallel corpora*

Beatrice Savoldi[1], Luisa Bentivogli[2]
Università degli Studi di Trento[1], Fondazione Bruno Kessler[2]
beatrice.savoldi@unitn.it, bentivo@fbk.eu

Recent years attested significant advancements in the quality of Machine Translation (MT). Such improvement is ascribed to the advent of neural networks, whose strength relies on their ability to uncover patterns and associations from the data they are fed with. In the case of MT, systems just need large corpora of parallel sentences to crunch. However, as cultural and societal biases enter their training data, MT models end up assimilating them, gender bias included.

Translation studies are acquainted with gender being sensitive information in cross-lingual transfer. The challenge ensues from structural asymmetries, namely how languages formally express human referents' gender via masculine or feminine markings (Corbett, 1991; Hellinger and Bußmann, 2001). In lack of any disambiguating information, translators facing the rendering of a gender-neutral word into a marked one make a personal choice, which reflects their own social gender assumptions (Nissen, 2002). However, this is not the case in MT.

Studies on the topic (Prates et al., 2018; Escudé Font and Costa-jussà, 2019) exhibited that MT models reproduce stereotypical gender roles (Romaine,1999; Talbot, 2003). Accordingly, *doctor* is inferred as a man, but *nurse* as a woman (Stanovsky et al., 2019), regardless of explicit cues about the referent's gender. Although such works represent valuable proxy estimations of gender bias, they only inspect gender assignment in a restricted domain, focusing on few highly connoted occupational nouns. However, grammatical-gendered languages like French, Italian, and Spanish extensively express gender via morphology on numerous parts of speech (Hockett, 1958). Moreover, MT deficit in translating gender is not restricted to stereotypical associations. Rather, Vanmassenhove et al. (2018) found that MT generally opts for masculine over feminine gender forms.

Prompted by the rising concern over biases in Natural Language Processing (Hovy and Spruit, 2016; Bender and Friedman, 2018; Savoldi et al., 2021), this study investigates gender translation from English into French, Italian, and Spanish. We believe that, as automatic models learn how to match cross-lingual features from linguistic data, fine-grained analyses on the phenomenon should start from the data themselves.

To this end, we conducted a corpus-based empirical study on 2500 sentences randomly sampled from MuST-C (Cattoni et al., 2020). Currently representing the largest multilingual corpus for Speech Translation (ST), MuST-C comprises sentence-aligned <audio, transcript, translation> triplets extracted from TED Talks. Focusing on its textual portion, the intent of our analysis was twofold: *i)* identify the circumstances in which referential gender assignment is implied in translation; *ii)* inspect how gender is realized and distributed across our three language pairs.

Accordingly, our cross-lingual manual analysis captured parallel sentences which, for human referents, require the translation of gender-neutral words into feminine or masculine marked ones. Their examination led to the classification of 4 distinctive circumstances for gender translation. Here, we introduce them through examples in English-Italian.

1. Like **my PhD advisor** *Revi* Sterling, *she,* of the **magic research high tops**…
   Come **la mia** tutor per il dottorato: Revi Sterling, che è **una delle** migliori **ricercatrici**…

The gender information is found within the sentence, i.e. the proper noun (*Revi*) and personal pronoun (*she*) inform about the *advisor*'s gender.

2. As **an** artist, connection is very important to me.
   Per **un'**artista come me, i legami sono fondamentali.

No information is available within the sentence. Since *an artist* refers to the speaker, proper feminine translation is only feasible when the speaker's gender is known.

3. *Camilla and I* have **been** to other organizations...
   Camilla e io siamo **state** in altre organizzazioni...

The inflection (feminine and plural) reflects the gender of both the speaker and a second participant (*Camilla*). To correctly assign gender, it is necessary to both know the speaker's gender and retrieve the contextual gender information for the second participant.

4. What do you think a batting average for a **cardiac surgeon** or a **nurse practitioner** or an **orthopedic surgeon**, **an OBGYN** is supposed to be?
   Quale credete debba essere la media di battuta per **un cardiochirurgo** o **una infermiera** professionista o **un chirurgo ortopedico**, **un'ostetrica**?

No information about the referents is available, gender assignment is arbitrary (and, as the human translation shows, potentially influenced by social expectations).

This 4-tiered categorization represents a useful scaffolding for future research, and it drove us into the creation of a subcorpus of MuST-C, primarily designed for the assessment of ST and MT quality in the translation of gender. It comprises around 1,000 <audio, transcript, translation> triplets for each of the three target languages addressed, with a subset of 450 common triplets for cross-lingual evaluations. The above-mentioned categories are represented in the corpus. The distribution of masculine/feminine gender-marked words is balanced across and within each language pair.

In our talk, we rely on the cross-lingual analyses carried out on our corpus to discuss the challenges, limits, and implications of gender translation in MT, ST, and human translation.

Moreover, as our examples show, gender markings occur on a great variety of parts of speech (POS): nouns, verbs, determiners, adjectives. This is relevant to fully understand the phenomenon of gender translation, especially for French, Italian and Spanish, languages with epicene nouns that - alone - do not convey gender distinction (see example 2: "artista"). In the interest of our analysis, we specifically isolated each gender-marked word in our corpus and inspected the proportion of gendered function words. Preliminary results show that articles (and articulated prepositions for Italian and Spanish) make up about 20% of all gender-marked expressions in the corpus. However, such distribution is uneven across feminine and masculine forms. To systematically analyze such dyssimetries and extend our study we enriched the multilingual corpus with an additional annotation layer concerning the POS of each gender-marked word. In our talk, we will discuss our findings from a contrastive perspective.

In the light of the above, the availability of our corpus is a valuable resource that can be used for fine-grained evaluations of automatic systems and to foster research on cross-lingual investigations. Thus, we

believe this study has significant implications for both the issue of gender bias in MT and the field of corpus linguistics.

References

Bender, E. M., & Batya Friedman. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. In *Transactions of the Association for Computational Linguistics. 6, 587-604.*

Corbett, G. G. (1991). *Gender.* Cambridge: Cambridge University Press.

Cattoni, R., Di Gangi M. A., Bentivogli, L., Negri, M., Turchi., M. (2020). MuST-C: A multilingual corpus for end-to-end speech translation. In *Computer Speech & Language Journal*. Doi: https://doi.org/10.1016/j.csl.2020.101155

Font, J. E., & Costa-Jussa, M. R. (2019, July-August). Equalizing gender biases in neural machine translation with word embeddings techniques. In *Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing.* ACL, Florence, Italy (147-154). Association for Computational Linguistics.

Hellinger, M., & Bußmann, H. (2001). Gender across languages: The Linguistic Representation of Men and Women. In M. Hellinger & H. Bußmann (Ed.) *Gender across Languages* (Vol. 2, pp. 1-25). Philadelphia, PA: John Benjamins Publishing Company.

Hockett, C. F. (1958). *A Course in Modern Linguistics.* New York, NY: Macmillan.

Hovy, D., & Spruit, S. L. (2016, August). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.* ACL, Berlin, Germany (591-598). Association for Computational Linguistics.

Nissen, U. (2013). Aspects of translating gender. *Linguistik Online*, *11*(2).

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2019). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, *1-19*.

Romaine, S. (1999). *Communicating Gender*. London: Lawrence Erlbaum Associates.

Savoldi, B., Gaido, M., Bentivogli, L., Negri., M., Turchi, M. (2021). Gender Bias in Machine Translation. To appear in *Transactions of the Association for Computational Linguistics.* MIT Press.

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019, July-August). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL. Florence, Italy (1679--1684). Association for Computational Linguistics.

Talbot, M. (2003). Gender Stereotypes: Reproduction and challenge. In J. Holmes & M. Meyerhoff (Ed.) *The Handbook of Language and Gender*. (pp. 468-86). Oxford: Blackwell Publishing.

Vanmassenhove,E., Hardmeier, C. & Way, A. (2018, October-November). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP, Brussels, Belgium (3003–3008). Association for Computational Linguistics.

But Are They Really the Same? A Contrastive and Parallel Study of French *mais* and Dutch *maar*

Nathanaël Stilmant, Gudrun Vanderbauwhede, Hanne Cardoen
Université de Mons, Faculté de traduction et d'interprétation, Ecole d'interprètes internationaux (FTI-EII)
Nathanael.STILMANT@umons.ac.be, Gudrun.VANDERBAUWHEDE@umons.ac.be,
Hanne.CARDOEN@umons.ac.be

Situated at the crossroads of contrastive linguistics and translation studies, our research focuses on the connectives of contrast in French and Dutch in the light of corpus analysis. Using a semasiological approach, we study a selection of the most representative connective of this category in two languages: *mais* (in French) and *maar* (in Dutch), the "quintessential contrastive [...] markers" (Fraser & Malamud-Makowski, 1996). Our criteria for the analysis of these connectives are based on the numerous studies that have analysed these connectives, which we have summarised in 5 axes. The first one is the semantic axis, which distinguishes *mais/maar* PA (translated as "pero" in Spanish and "aber" in German), *mais/maar* SN ("sino" in Spanish and "sondern" in German) (Anscombre & Ducrot, 1977) and a *mais/maar* of semantical opposition (Van de Voorde, 1992; Haeseryn, 1997; Giacalone & Camugli, 2011), which we will call *mais SO,* as the three fundamental categories within which we find five main distinctions (concessive, adversative, additive, phatic, and narrative *mais/maar*) as well as independent minor categories, like interrogation or surprise (Gettrup & Nølke, 1984; Adam, 1990; Pelletier, 1992; Lamiroy & Van Belle, 1995; Rabatel, 1999; Uusialho, 2000; Bacha, 2005; Pelizzoni, 2009). The indirectness or directness of *mais* is also investigated (Moeschler & De Spengler, 1982), as well as a possible "denial of expectation" in the proposition following *mais/maar* (Lakoff, 1971; Kerbrat-Orecchioni, 1999). Our second axis of study focuses on the grammatical nature of the elements joined by *mais/maar* (Birkelund, 2009). The place of the connective in the sentence constitutes the third axis (beginning, middle, or end of the sentence), the fourth axis deals with the length of the elements linked by *mais/maar*, and the fifth axis with collocations and connective sequences (Luscher, 1993; Razgouliaeva, 2010). These criteria are intended to be as broad as possible so as not to neglect any aspect of research. Having organised these criteria, we carried out two different corpus studies: two twin monolingual studies, one of 100 occurrences of *mais* (journalistic corpus from the Est Républicain, Université Toulouse 2), the other of 100 occurrences of *maar* (SoNaR corpus, Nederlandse Taalunie); and a parallel study of 100 occurrences of *maar*/*mais* (Dutch Parallel Corpus, Katholieke Universiteit Leuven). The monolingual studies investigated the relative frequencies of the above-mentioned categories, described the use of *mais* and *maar* in their respective languages using our criterion set, and compared the uses of *mais* with those of *maar*. The parallel study investigated the different ways in which professional translators translate *maar* and whether the different translation processes observed (literal translation of *maar* by *mais*, absence of connective or choice of another connective than *mais* in French) vary according to the category to which *maar* belongs. The results of the two monolingual studies were then combined with those of the bilingual analysis and analysed according to the principles of inferential statistics using the R programme. Monolingual studies have shown the distributional properties of *mais* and *maar*. We observed similarities between the two languages, such as a similar distribution of *mais* SN and *maar* SN (about 5% of the corpora), and differences, like the fact that *mais* SO is found more often in French than *maar* SO in Dutch (8% and 2% of the corpora respectively). The subcategories also show differences in distribution: in French, addition *mais* are the most numerous, followed by adversation, concession, narration and phatic *mais*. In Dutch, the most common *maar* are concessive, followed by adversative, additive, narrative and phatic *maar*. Besides these differences, we also found similarities between French and Dutch. In both languages, the concessive occurrences of *mais*/*maar* are mostly direct (even if the directness is more dominant in French than in Dutch) and in each language, it is the direct concessive relations that present the highest number of "denial of expectation". On the other hand, adversation and addition *maar*/*mais* are overwhelmingly

indirect in both French and Dutch. The study of the length of the elements joined by *mais*/*maar* also revealed interesting properties. Again, there are similarities and differences. In the majority of cases, sentences with direct relationships are on average longer than those with indirect relationships, in both French and Dutch, for all semantical subcategories to which the directness or indirectness applies (concession, adversation, and addition). On the other hand, narrative *mais* are the longest in French, while the phatic *maar* are the longest in Dutch. Then, both languages observe the same length ranking (addition, concession, and finally adversation). The parallel study showed that the cases of literal translation, without any modification (of notion, grammatical nature, place or collocation) are about half of the cases of *maar* SO and *maar* SN, but only less than one third of the cases of *maar* PA. Concession is the best represented category in the unmodified translations, followed by narrative, additive, adversative and phatic *maar*. Among the modifications observed, 21 *mais* PA are translated by something other than *mais*, i.e. 25% of the occurrences. Half of them are translated by an "unmarked relationship" (Corminboeuf, 2014), the other half by another connective which generally expresses more typically in French the relationship established in the original Dutch sentence (in this case it is mostly a concession, then an adversation and finally an addition, but never a narrative or phatic *maar*). The new connectives used in such case are various (*cependant, et, même si, or, par contre, pourtant, si*). They show the great polysemy that *mais/maar* can have and the large number of notions they can be used for. *Maar* SN and *maar* SO are more often translated literally (only one occurrence of unmarked relationship in the corpus for each of these categories, and no cases of connective change). Although *mais* and *maar* form a pair of connectives, they do not have exactly the same properties in their respective languages. Studying these properties can help us to understand better how they are translated.

References

Adam, J-M. (1990). *Eléments de linguistique textuelle*. Liège: Mardaga.

Anscombre, J-C. & Ducrot, O. (1977). Deux mais en français ? *Lingua*, *43* (1), 23-40.

Bacha, J. (2005). Le fonctionnement de *mais* dans l'*Heptaméron* de Marguerite de Navarre. *L'information grammaticale, 107*, 57-60.

Birkelund, M. (2009). Pierre n'est pas français mais danois. Une structure polyphonique à part. In : *Langue Française*, *4*(164), 123-135.

Corminboeuf, G. (2014). *L'identification des relations de discours implicites : le cas de l'adversation.* Paper presented at Congrès Mondial de Linguistique Française, Berlin.

Fraser B & Malamud-Makowski M. (1996). English and Spanish contrastive discourse markers. *Language Siences, 18*, (3). 863-881.

Gettrup, H. & Nølke, H. (1984). Stratégies concessive : une étude de six adverbes français. *Revue romane, 19*, 3-47.

Giacalone Ramat, A., Camugli Gallardo, C. (2011). L'emploi des connecteurs : però correspond-il toujours à mais ?. *Revue française de linguistique appliquée*, *16*, 57-74.

Haeseryn, W. (1997). *Algemene Nederlandse Spraakkunst*. Deurne: Martinus Nijhoff Uitgevers.

Kerbrat-Orecchioni, C. (1999). *L'énonciation* (4th ed.). Paris: Armand Colin.

Lakoff, R. (1971). If's, and's, and but's about conjunction. In C. Fillmore & D. Langendoen (Eds.), *Studies in Linguistics Semantics* (pp. 115-150). New-York: Reinhart and Wilson

Lamiroy, B & Van Belle, W. (1995). Connectives of contrast and concession in Dutch and French. *Leuvense bijdragen, 84(3)*, 387-418.

Luscher, J.-M. (1993). La marque de connexion complexe. *Cahiers de linguistique française*, *14*, 173-188.

Moeschler, J. & De Spengler, N. (1982). La concession ou la réfutation interdite, approches argumentative et conversationnelle. *Cahiers de linguistique française, 4*, 7-36.

Pelizzoni, F. (2009). *Le connecteur pragmatique "mais", son rôle et sa fonction dans les séquences dialoguées de trois pièces de théâtre (Zaïre, Hermani, Les bonnes)* (Master's thesis, Università degli studi di Verona).

Pelletier, C. (1992). *Étude des connecteurs ET et MAIS dans des productions écrites d'étudiants universitaires : approche sémantico-pragmatique* (Master's thesis, Université du Québec à Chicoutimi).

Rabatel, A. (1999). *Mais* dans les énoncés narratifs, un embrayeur du point de vue et un organisateur textuel. *Le français moderne, 67(1)*, 49-60.

Razgouliaeva, A. (2010). Combinaison des connecteurs *mais enfin*. *Cahiers de linguistique française*, *24*. 143-168.

Uusialho, O. (2000). *Le* mais *argumentatif dans trois romans contemporains*. (Master's thesis, Jyväskylän yliopisto).

Van de Voorde, K. (1992). De deux à trois "mais" : essai de vérification des approches d'Anscombre et Ducrot et de Blumenthal. *Travaux de Linguistique, 24*, 57-81.

Using parallel corpora for researching patterns of grammatical variation: the case of nominal word order in European languages

Luigi Talamo
Saarland University
luigi.talamo@uni-saarland.de

Cross-linguistic studies often provide us with an over-simplistic representation of grammatical structures, forcing grammatical variation into discrete categories and giving rise to what has been addressed by Wälchli (2009) as the 'bimodal distribution bias'. For instance, map no. 87A of the World Atlas of Linguistic Structure, 'Order of Adjective and Noun' (Dryer 2013), shows us the picture of a linguistic Europe split into two parts: a southwestern part with Noun-Adjective languages, corresponding to Romance and Celtic languages, and a northeastern part with the reversed order, featuring the remaining European languages, with the isogloss traversing the centre of the *Charlemagne Sprachbund*.

However, this neat picture is readily falsified by the actual linguistic usage; for example, in French and Italian the position of adjectives may be influenced by functional/pragmatic factors, as well as by semantics (Price 2013:196-207, Maiden & Robustelli 2013:48-57), and some European languages, such as Polish, do not actually show a dominant Adjective-Noun order (Swan 2002:127-129). Furthermore, traditional data sources for contrastive and typological studies (grammars) do not always offer a thorough representation of the variation observed in actual linguistic usage, especially for poorly studied grammatical structures and/or languages with few and outdated descriptions.

In order to fill the gap between actual linguistic usage and typological studies, hence overcoming the bimodal distribution bias, it has been suggested to treat linguistic data as continuous data-points, which are sourced from input such as visual stimuli, judgement tasks and corpora, and are handled using probabilistic methods. This type of approach has been applied to the study of different grammatical structures, including causatives (Levshina 2015), motion verbs (Verkerk 2014), contrastive negation (Silvennoinen 2020) and word order (Levshina 2019), and is aptly named 'token-based typology' (Haspelmath 2019).

One of the best data sources for token-based studies is represented by parallel corpora, which are however scarcely available and are biased towards religious, technical or legal contents (Christodoulopoulos & Steedman 2014, Agić & Vulić 2019, Tiedemann 2012). As its name suggests, the parallel Corpus of modern Indo-European Prose (CIEP: Talamo & Verkerk: submitted) aims to represent a different genre, fiction, which is closer to spoken varieties; at the time of writing, CIEP features 18 books in 15 languages from five different Indo-European genera (Balto-Slavonic, Celtic, Germanic, Greek, Romance). The corpus has been automatically annotated for lemma, parts of speech (both universal and language-specific) and syntactic relations using a parser trained on Universal Dependency models (Zeman et al. 2019).

Our contribution explores in the CIEP corpus the word order variation of 6 adnominal modifiers: article, demonstrative, locative analytic case marker (adposition), modifying adjective, quantifier and relative clause; these modifiers are formulated as comparative concepts i.e., "concepts specifically designed for the purpose of comparison" (Haspelmath 2010:666) and are matched with different layers of UD annotation:

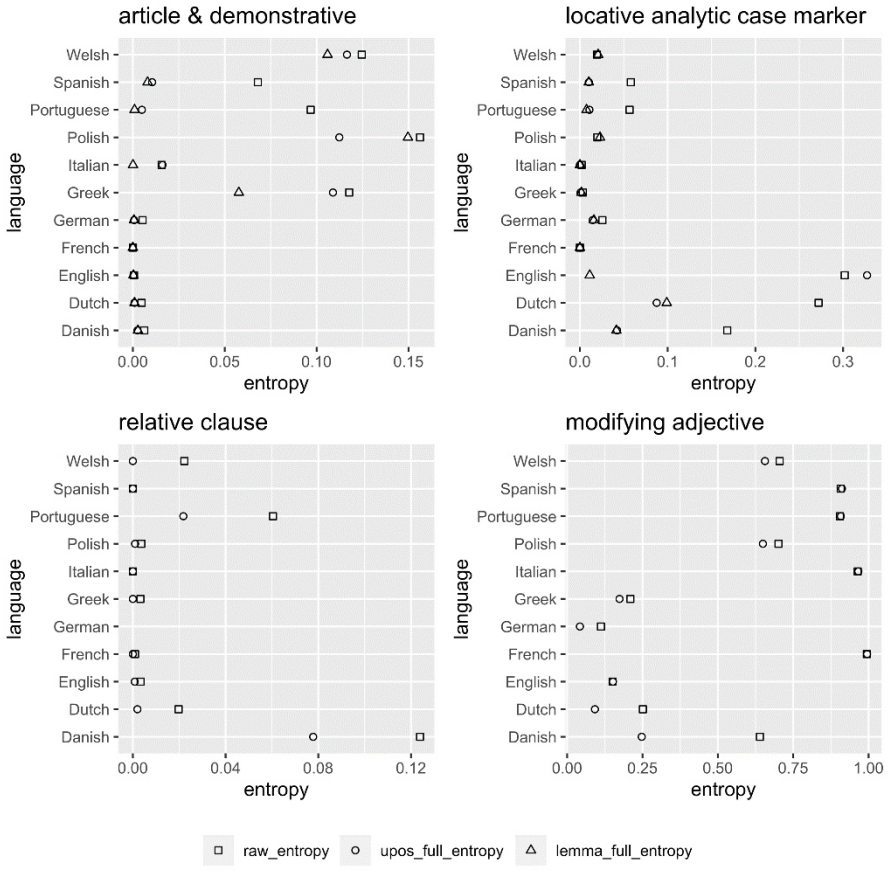
- the syntactic layer, which makes use of relevant UD relations such as determiner, numeral, case marking, adjectival modification and relative clause modifier;

- the parts-of-speech layer, which refines the syntactic layer by including Universal part of speech labels such as the noun and proper noun tags for nominal heads and the adjective tag for modifying adjective;
- the lexical layer, which consists of language-specific list of lemmas that are sourced from the corpus and descriptive grammars, and are particularly effective in identifying adnominal relations relying on closed word classes such as the article, demonstrative, locative analytic case and quantifier.

These layers can be applied alone or combined together in order to analyze corpus data through different filters:

1. a 'raw' filter, which uses only the syntactic layer;
2. a 'parts-of-speech' filter, which combines the parts-of-speech layer with the syntactic layer;
3. a 'lemma filter', which adds lexical information to the 'parts-of-speech' filter;

We express word order variability in terms of Shannon's entropy (Shannon 1948), an information theoretic measure that is based on probability distribution. The distribution of word order has two possible variables: head-modifier and modifier/head; with a 50%-50% probability distribution, the entropy rate reaches its maximum value of 1; with no variation i.e., just one of the possible orders, the entropy rate is at its minimum value of 0.



Our preliminary results, which are restricted to 11 languages and 5 comparative concepts, challenge the categorical representation of traditional typological studies both at the cross-linguistic and language-specific level (see the figure above). The raw filter (raw_entropy) captures high rates of entropy (>0.6-0.7) for modifying adjectives in Romance, Celtic and Balto-Slavonic languages, however confirming very low variability for article, demonstrative and relative clauses in all languages. The parts-of-speech filter (upos_full_entropy) validates some unexpected patterns of variation observed in the raw filter, such as

relative sensible rates of entropy for modifying adjectives in Danish and Greek (~0.22), and articles/determiners in Greek and Polish (~0.13); finally, the lemma filter (lemma_full_entropy) gives reliable evidence for the variability of locative analytic case marker in Germanic languages such as English (~0.3) and, to a minor extent, Dutch and Danish (~0.15-0.25).

References

Agić, Ž. & Vulić, I. (2019). JW300: A Wide-Coverage Parallel Corpus for Low- Resource Languages. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3204–3210. doi: 10.18653/v1/P19-1310. url: https://www.aclweb.org/anthology/P19-1310.

Christodoulopoulos, Ch. & Steedman, M. (2014). A massively parallel corpus: the Bible in 100 languages. In: *Journal of Language Resources and Evaluation* 49.2, pp. 375–395.

Dryer, M. (2013). Order of Adjective and Noun. In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. url: https://wals.info/chapter/87.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in cross- linguistic studies. In*: Language* 86.4.

Haspelmath, M. (2019). How comparative concepts and descriptive linguistic categories are different. In: *Aspects of Linguistic Variation*. Ed. by Tanja Mortelmans Daniël Van Olmen and Frank Brisard. Berlin: Walter de Gruyter.

Levshina, N. (2015). Lexical Typology through Similarity Semantics: Toward a Semantic Map of Motion Verbs. In: *Folia Linguistica* 49.2, pp. 487–520. url: https: //doi.org/10.1515/ling-2012-0021.

Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. In: *Linguistic Typology* 23.3, pp. 533–572.

Maiden, M. & Robustelli. C. (2013). *A reference grammar of modern Italian*. 2nd ed. London & New York: Routledge.

Price, G. (2013). *A Comprehensive French Grammar*. 6th ed. Oxford: Blackwell, p. 610.

Shannon, C. (1948). A mathematical theory of communication. In: *Bell Syst. Tech. J.* 27.3, pp. 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x. url: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Silvennoinen, O. (2020) Comparing corrective constructions: Contrastive negation in parallel and monolingual data. In: *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Ed. by Tanja Mortelmans Daniël Van Olmen and Frank Brisard. Berlin: Walter de Gruyter, pp. 221–264.

Swan, O. (2002). *A grammar of contemporary Polish*. Bloomington: Slavica, p. 496.

Talamo, L. & Verkerk A. Nominal heads and modifiers in modern European languages: a token-based typology of word order. Submitted.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12). Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA). isbn: 978-2-9517408-7-7.

Verkerk, A. (2014). Diachronic change in Indo-European motion event encoding. In: *Journal of Historical Linguistics* 4.1, pp. 40–83.

Wälchli, B. (2009). Data Reduction Typology and the Bimodal Distribution Bias. In: *Linguistic Typology* 13.1, pp. 77–94.

Zeman, D. et al. (2019). Universal Dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. url: http://hdl.handle.net/11234/1-3105.

Torn between source language constructions and target language expectations. Translating passive construal

Isabel Tello[1], Ulrike Oster[2]
Universidad Politécnia de Cartagena[1], Universitat Jaume I[2]
isabel.tello@cud.upct.es, oster@uji.es

For the linguistic description of an event, the speaker chooses which perspective to adopt, which aspects to foreground and which constructions to use. In Cognitive Linguistics terms, the speaker's construal (Langacker 1987) of the event determines how (s)he will communicate it linguistically. One fundamental factor of event construal is the relative importance assigned to each participant (agent, patient, action or process). For example, whether the action is conceptualised from an active or a passive perspective.

This paper concentrates on the passive construal of events from an interlingual point of view. For this purpose, *passive construal* is defined as one in which the agent is defocused (Agensausblendung) while the action and (optionally) the patient take centre stage. Such a construal can be expressed through various constructions in the languages considered in this study (English, French, German and Spanish). The most prominent one in English is undoubtedly the passive voice ("He was found guilty"), alongside less frequent structures like middle sentences ("This book reads easily") (Leech & Svartvik 1975). French and German behave similarly. However, both offer a greater variety of alternative constructions ("on", "se", "man", "bekommen + PP", etc.) while the passive voice in the strict sense ("être/werden + PP") is less predominant (Chevalier 1995; Cartagena & Gauger 1989). Spanish, on the other hand, is to be found on the other extreme of the continuum: the passive voice ("ser + PP") is considered less prominent than other structures. These include the reflexive passive ("Se abren las puertas"), the impersonal passive ("Se habla inglés") or the impersonal active using the 3rd person in the plural ("Lo tiraron al suelo") (Lavid et al. 2010; Valero Garcés 2012).

However, the study addresses passive construal not just from a contrastive linguistic but from a translational angle. The paper's main objective is thus to describe translator behaviour when it comes to rendering constructions that express passive construal from three different languages (English, German and French) into Spanish.

The theoretical framework is provided by Halverson's (2003, 2010, 2017) *Gravitational Pull Hypothesis* (GPH). GPH proposes a cognitive basis for universal features of translation (Baker 1993). One of the most widely researched of these features is normalisation or standardisation, which is believed to lead to an overrepresentation of target language-specific features. Evidence for this hypothesis has been collected across several language combinations and for different linguistic structures. Conversely, Tirkkonen-Condit's *Unique Items Hypothesis* (2004) predicts underrepresentation of target-language (TL) items with no counterpart in the source language (SL). GPH explains this apparent contradiction invoking the cognitive mechanisms of the bilingual mind and suggests that both over and underrepresentation of particular TL items are possible (Halverson, 2010: 352). Depending on the characteristics of SL and TL, GPH predicts the outcome of a given translation situation according to three factors (Halverson 2010: 356): "patterns of prototypicality in the target language", leading to overrepresentation in the translation, "conceptual structures or the representation of the source language item", also leading to overrepresentation, and "patterns of connectivity", which reflect relationships between source language and target language items and could lead to over or underrepresentation.

Regarding passive construal, the following structures will be considered:

- From a SL perspective (EN, FR, DE), the passive in a narrow sense ("to be/werden/être + PP").

- From a TL perspective, the three main Spanish constructions expressing passive construal: reflexive and impersonal passive ("se"-constructions), the passive ("ser + PP") and the impersonal active (3rd p. pl.).

The following (potentially conflicting) tendencies are hypothesised:

A. Regarding "ser + PP": **Overrepresentation**
   Since the passive constructions "to be/werden/être + PP" are salient in English, French and German, the corresponding Spanish structure will be overrepresented in Spanish translations (gravitational pull). It is further hypothesised that this effect is especially noticeable in English (due to the marked salience of "to be + PP") but less so in French and German.
B. Regarding the impersonal 3rd person plural: **Underrepresentation**
   Since this structure can be considered a unique item for Spanish with respect to all three source languages, it will be underrepresented in translations.
C. Regarding the reflexive/impersonal passive ("se"-constructions):
   Over- or underrepresentation due to potentially **conflicting tendencies**:
   - It is a salient structure in Spanish, which might lead to an **overrepresentation** in Spanish translations (magnetism of TL structures).
   - On the other hand, "se"-constructions are a unique item in the TL in relation to English and German, possibly leading to its **underrepresentation** in translations from these languages.

Methodologically, the study draws on Halverson (2017) Hareide (2017a, 2017b), Marco & Oster (2018), Oster (2020) and Marco (2019, 2021). The hypotheses were tested through the analysis of the relevant structures in the multilingual, parallel and comparable COVALT corpus. This corpus contains narrative texts originally written in English, French and German and their translations into Catalan and Spanish plus two subcorpora of Catalan and Spanish comparable original works.

Phase I. Analysis of COVALT as a comparable corpus

- Research question A: Which structures are most salient in each language?

After a preliminary step in which the constructions matching our definition of *passive construal of an event* were identified in all four languages, a frequency analysis of each of these linguistic structures established their categorical salience in original texts (EN, FR, DE, ES).

- Research question B: How does translated Spanish differ from original Spanish?

The same analysis was applied to Spanish texts translated from English, French and German. Results were then compared to those of original Spanish.

Phase II. Analysis of COVALT as a parallel corpus (EN-ES, FR-ES, DE-ES)

- Research question C: What are the connectivity patterns between source and target language items, and how strong are these links in terms of target and source concentration (cf. Halverson 2017)?

This required an analysis of **Spanish translation solutions** for English, French and German passive ("to be/werden/être + PP"), on the one hand, and the **triggers** in English, French and German source texts for the Spanish passive ("ser + PP"), on the other.

Results from Phase I confirm that passive constructions in general as well as the passive voice in a strict sense are much less frequent in Spanish than in English, French and German. As expected, Spanish "ser + PP" is also less frequent than other Spanish passive constructions like the impersonal/reflexive passive or the impersonal 3rd p. pl. As to the comparison of original and translated Spanish, hypothesis A (overrepresentation of "ser" + PP) could be confirmed for EN-ES and FR-ES. This is consistent with the gravitational pull of the prototypical passive construction in English and French, leading to an overuse of the corresponding Spanish structure. In translations from German, however, in which there is no clear predominance of "werden + PP", this tendency is overridden by the magnetism (cf. Halverson 2017) of

other salient TL structures. Hypothesis B (underrepresentation of 3rd p. pl.) was confirmed for all three language combinations. As to hypothesis C, overrepresentation of "se"-constructions was found for DE-ES and EN-ES but not for FR-ES. This seems to indicate that the uniqueness of an item is of minor importance in the case of salient structures.

Additionally, the analysis of patterns of connectivity in Phase II provided detailed insights into how SL and TL structures are connected in each translation pair and how strong their links are. It also led to the identification of language pair-specific "preferred translation routes".

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In F. Gill, M. Baker & E. Tognini-Bonelli (eds.) *Text and Technology. In Honour of John Sinclair.* Amsterdam & Philadelphia: John Benjamins, 233-250.

Cartagena, N., & Gauger, H. M. (1989). *Vergleichende Grammatik Spanisch-Deutsch*. Mannheim: Dudenverlag.

Chevalier, J. C., Arrivé, M., Blanche-Benveniste, C. & Peytard, J. (1995). *Grammaire Larousse du français contemporain*. Paris: Larousse.

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

Halverson, S. (2010). Cognitive translation studies: Development in theory and method. In G. M. Shreve & E. Angelone (eds.) *Translation and Cognition*. Amsterdam & Philadelphia: John Benjamins, 349-369.

Halverson, S. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M. A. Lefer & I. Delaere (eds.) *Empirical Translation Studies*. Berlin: De Gruyter Mouton, 9-45.

Hareide, L. (2017a). The translation of formal source-language lacunas. An empirical study of the Over-representation of Target-Language Specific Features and the Unique Items Hypothesis. In M. Ji, M. Oakes, L. Defeng, & L. Hareide (eds.) *Corpus Methodologies Explained. An Empirical Approach to Translation Studies*. London & New York: Routledge, 137-187.

Hareide, L. (2017b). 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, & Lidun Hareide (eds.) *Corpus Methodologies Explained. An Empirical Approach to Translation Studies*. London and New York: Routledge, 188-231.

Langacker, R. (1987). *Foundations of Cognitive Grammar. I. Theoretical Prerequisites (Vol. 1)*. Stanford: Stanford University Press.

Lavid, J., Arús, J. & Zamorano-Mansilla, J. R. (2010). *Systemic Functional Grammar of Spanish*. *A Contrastive Study with English*. London & New York: Continuum.

Leech, G & Svartvik, J. (1975). *A Communicative Grammar of English*. London: Routledge.

Marco, J. (2019). Living with parallel corpora. The potentials and limitations of their use in translation research. In I. Doval Reixa & M. T. Sánchez Nieto (eds.) *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*. Amsterdam & Philadelphia: John Benjamins, 39-56.

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*. Berlin: Language Science Press. 27–52.

Marco, J. & Oster, U. (12-14 September 2018). *The gravitational pull of diminutives in Catalan translated and non-translated texts*. Using Corpora in Contrastive and Translation Studies Conference (5th edition). Louvain-la-Neuve, Belgium.

Oster, U. (2020). "Sobrerrepresentación del adjetivo antepuesto en textos traducidos: ¿realidad o prejuicio?". En Recio Ariza, M.ª Ángeles et al. (eds). *Del texto a la traducción. Estudios en homenaje a Pilar Elena*. Granada: Comares. 115-132.

Tirkkonen-Condit, S. (2004). Unique items – over – or under-represented in translated language? In A. Mauranen & P. Kujamäki (eds.) *Translation Universals: Do They Exist?* Amsterdam & Philadelphia: John Benjamins, 177-184.

Valero Garcés, C. (2012). Inglés y español mano a mano: Dos lenguas y dos formas de ver el mundo. Cuadernos Cervantes, Época II – Año III. http://www.cuadernoscervantes.com/lc_ingles.html Accessed 25 February 2021.

Distributional Lexicon in Contrast

Aleksandar Trklja
University of Innsbruck
aleksandar.trklja@uibk.ac.at

According to the dominant view in cognitive linguistics lexical meaning is conceptual (e.g. Jackendoff, 1983; Pinker and Levin, 1991; Johnson and Lakoff, 2002). Concepts are considered to be the building blocks of meaning. There is no consensus as to what concepts are but it is generally assumed that they stand for clusters of some sorts of primitive semantic units (e.g. Jackendoff, 1983; Barsalou, 1992). This approach to lexical semantics suffers from several shortcomings (e.g. Dixon, 1971; Van Roey, 1990). First, it is far from clear what the ontological nature of semantic primitives is. As Gordon (2003: 2219) notes, "[t]here is nothing to suggest the existence of any objective or universally applicable means of establishing parameters for a componential analysis". It is no accident that componential studies are usually limited to the study of nouns that refer to concrete objects or verbs that denote physical activities such as motion verbs. It is relatively easy to identify semantic borders for such words "[b]ut many other vocabulary terms refer to 'things' which have features that are not neatly distinguishable, so that their meanings have 'fuzzy edges', i.e. contrast only vaguely and cannot be adequately described in terms of components" (Van Roey, 1990: 30). One of the serious challenges for contrastive studies is how to match semantic primitives between languages. Although semantic labels and meaning components are based on *a priori* established categories the criteria that underlie these categories are not universal across languages. In fact, cognitive psychologists hold that frames and concepts "are continually updated and modified due to ongoing human experience" (Evans and Green, 2006: 223). Nevertheless, in linguistic studies the categorise which are established in one language (typically English) are applied to other languages. For example, the theory of frame semantics was initially developed only for English and was subsequently applied to other languages (e.g. Boas, 2005; Braasch, 1994). It means that the frames applied to other languages are adopted from English. As a consequence, the frames that do not exist in English but do exist in other languages remain invisible here. Any approach that relies solely on such pre-established categories, therefore, cannot provide a comprehensive description of lexical semantics across languages. Finally, approaches based on conceptual analysis provide no explanation of how to identify corresponding items from two or more languages.

In the present paper, it will be argued that a more comprehensive account of the contrastive lexicon can be achieved through a three-stage approach that combines distributional corpus methods with event semantics. The approach is based on the assumption that "the categorization mechanism of the language learner [= speaker] is driven by the superficial distributional properties of the data that is dealing with" (Culicover, 1999: 85). I will argue that the learner's semantic categorization is based partly on her familiarity with the occurrence of lexical items and partly on the properties of eventuality.

At the first stage, the lexical items from two or more languages are identified in a parallel corpus. Distributional properties of lexical items are defined in terms of the unique sets of cooccurrents (Harris, 1957; Sinclair, 1991; Croft, 2002). In the present model, the lexical items from two or more languages that share distributional properties are grouped into lexical domains or equivalence classes. Such classes are identified in parallel corpora following the extended distributional hypothesis (Trklja, 2017):

> No two items from one language will correspond to the same item from another language and simultaneously occur in the same context unless they have the same meaning.

All the items that satisfy this condition are regarded as members of the same equivalence class.

After that, the common distributional properties are investigated by means of local grammars (Gross, 1993; Hunston and Sinclair, 2000). Local grammars are based on "a purely wordcombinatorial investigation" (Harris, 1988: 40) instead of some pre-established categories. Local grammars are templates composed of *ad hoc* categories derived from the co-occurrence relations and they provide a fine-grained description of distributional properties. It will be argued that such properties are indicative of lexical structures and it will also be illustrated how the local grammar approach can be extended to the contrastive lexical semantic analysis. Figure 1 illustrates how shared distributional properties of the English lexical items that occur in the equivalence class called {CAUSE PROBLEM} can be identified in a local grammar and represented in terms of a finitestate automaton. In a similar manner, a local grammar of the corresponding items from other languages can be represented.

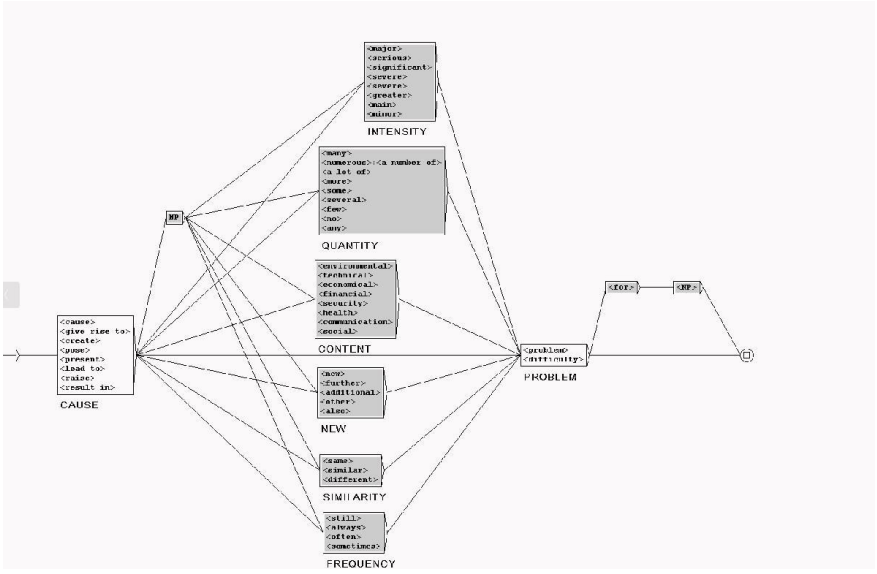


Figure 1: Local grammar diagram of {CAUSE PROBLEM}

Lexical items differ in terms of how many relations they are able to establish with the items from other languages and how strong those relations are. This has been modelled at the second stage in the present paper in terms of the notion of *correspondence degree*. Table 1 represents the lexical items from English and German with their values of correspondence degree.

Lexical items	Correspondence degree
<cause> *problems*	13.2
<there be> *problems*	12.4
<create> *problems*	11.8
problems <arise>	8.7
<give rise to> *problems*	8
<pose> *problems*	7.2
<present> *problems*	7
<raise> *problems*	4.7
<to be problematic>	4.5
<lead to> *problems*	4
<result in> *problems*	4

Lexical items	Correspondence degree
<zu> *Problemen* <führen>	12.1
Probleme <bringen>	10.4
Probleme <es gibt>	9.5
Probleme <verursachen>	8.4
Probleme <schaffen>	7.3
Probleme <bereiten>	7.3
Probleme <auftreten>	6.8
Probleme <entstehen>	6.6
Probleme <aufwerfen>	6.4
<problematisch sein>	5.7
Probleme <sich ergeben>	5.3
Probleme <darstellen>	4
<Ursache GEN|für> *Probleme*	3

Table 1: Lexical items from an English and German corresponding lexical domain

At the final stage, the differences between lexical items across languages are explored both in terms of fine-grained distributional properties and the properties of eventuality (e.g. Parsons, 1990; Rothstein, 2004). Distributional properties are modelled in terms of statistical association measures and regression analysis and the properties of eventuality are accounted in terms of the aspectual meaning and telicity.

The final model provides a comprehensive contrastive description of lexical meaning.

References

Barsalou, L.W. (1992) "Frames, Concepts, and Conceptual Fields." In: Lehrer, A. and Kittay E.F. (eds.) *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization.* London: Routledge, pp. 21-74.

Boas, H.C. (2005) "Semantic Frames as Interlingual Representations for Multilingual Lexical Databases." *International Journal of Lexicography*, 18 (4), pp. 445–478.

Braasch, A. (1994) "There's no Accounting for Taste - Except in Dictionaries." *Proceedings. Amsterdam: EURALEX.* Amsterdam, pp. 45-55.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Culicover, P. W. (1999) Syntactic Nuts, Oxford: Oxford University Press.

Dixon, R.M.W. (1971) "A Method of Semantic Description." In: Steinberg, D. and Jakobovits, L. (eds.) *Semantics*. Cambridge: Cambridge University, pp. 436–471.

Evans, V. and Green, M. (2006) *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Gordon, W.T. (2003) "Semantic Theories in 20th-century America: An Overview of Approaches Outside Generative Grammar" In: Wiegand, H.E. (ed.) *History of the Language Sciences.* Berlin: de Gruyter, pp. 2213-2229.

Gross, M. (1993) "Local Grammars and their Representation by Finite Automata." In: Hoey, M. (ed.) *Data, Description, Discourse. Papers on the English Language in Honour of John McH Sinclair*. London: Collins, pp. 26–38.

Harris, Z. S. (1957) "Co-occurrence and transformation in linguistic structure." *Language*, 33(3), 283-340.

Hunston, S. and Sinclair, J. M. (2000) "A Local Grammar of Evaluation." In: Hunston S. and Thomson, G. (eds.) *Evaluation in Text: Authorial Stance and the Construction of Discourse.* Oxford: Oxford University Press, pp. 74–101.

Jackendoff, R. (1983) *Semantics and cognition*. Cambridge, Mass.:MIT press.

Johnson, M., and Lakoff, G. (2002) "Why cognitive linguistics requires embodied realism." *Cognitive linguistics*, 13(3), 245-264.

Parsons, T. (1990) *Events in the Semantics of English*. MIT Press, Cambridge.

Pinker, S., and Levin, B. (1991) *Lexical and conceptual semantics*. Cambridge, Mass.:MIT press.

Rothstein, S. (2004) S*tructuring events: A study in the semantics of lexical aspect*. Malden, MA: Blackwell. Chicago

Sinclair, J.M. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Trklja, A. (2017) "Corresponding lexical domains: A new resource for onomasiological bilingual dictionaries." *International Journal of Lexicography* 30 (4), 416-453.

Van Roey, J. (1990) *French-English Contrastive Lexicology: An Introduction.* Leuven: Peeters Publishers.

How Revealing: The Embedded Exclamative in Translation

Faye Troughton
University of Mons
Faye.troughton@umons.ac.be

This paper reports on both a contrastive study of original *how* exclamatives in English (1–2) and original *combien* exclamatives in French (3–4) in spoken political discourse and a study of their translation into French and English respectively.

(1) *How right that is!* (SPEAKER ID=271 NAME=Jackson (PPE-DE))
(2) *They think it is just another piece of paper saying how wonderful they are.* (SPEAKER ID=204 NAME=Martin, David W.)
(3) *Je voudrais faire une première remarque: combien avons-nous été imprudents en renonçant aux instruments que nous donnait le traité CECA!* (SPEAKER ID=53 NAME=Herman)
(I would first of all point out how unwise we were to condemn the instruments that gave us the ECSC Treaty.)
(4) *Débat qui montre aussi combien il est difficile, aujourd'hui, de dissocier le traitement des questions sociales au plan communautaire, des débats politiques nationaux.* (SPEAKER ID=121 NAME=Xavier Bertrand)
(Our debate also shows how hard it is, nowadays, to separate the consideration of social issues at Community level from policy debates at national level.)

In exclamative constructions (1) and (2), *how* acts as a qualitative degree modifier, indicating the extremely high degree of rightness or wonderfulness that the speaker wishes to convey (cf. Quirk et al.1985: 834, Huddleston & Pullum 2002: 919). In (3) and (4), the French exclamative word *combien* acts in a similar way, expressing a high degree of unwiseness and difficulty respectively.

While there is a general consensus that the construction illustrated in (1) constitutes an exclamative, there is some division as to whether (2) can be defined as such. Some ascribe to the view, outlined in some reference grammars (Quirk et al 1985: 1055; Huddleston and Pullum 2002: 991), that exclamative clauses can be embedded in much the same way as interrogative clauses (c.f. Elliot 1974: 233; Michaelis & Lambrecht 1996; Michaelis 2001; Zanuttini & Portner 2003; Collins 2005). Others draw a clear distinction between clause-initial and embedded constructions, arguing that they cannot be included under the umbrella term "exclamative" and so define these constructions as complement or embedded *wh-* clauses (Heine et al. 2020; Rett 2008, 2011). According to Rett (2008: 603), the exclamative is encoded in terms of its illocutionary force and as this "is a property of an utterance as a whole, not subcomponents", the term "exclamative" may only apply to matrix clauses. Heine et al. (2020: 216) share this view arguing that embedded constructions take the "argument status" of the matrix clause, whereas true exclamatives are syntactically unattached and are not arguments.

Decisions made by translators when faced with these constructions may contribute to this discussion. Indeed, in a study of the translation of exclamatives using *what* in English and *quel* in French, Troughton (subm.) remarked that translators seemed to interpret these embedded constructions differently and were less likely to use an embedded exclamative in the target language. The present study, involving the only other English exclamative word, *how*, helps both to ascertain if this observation applies to exclamatives in general and demonstrate the contribution that corpus-based translation studies can make to debates on how linguistic phenomena are defined.

This study aims to answer the following questions.

a) How are *how* exclamatives in English and *combien* exclamatives in French used in political discourse?
b) How are these constructions translated in practice?
c) Do these translations have any implications for the status of the embedded exclamative?

These are answered through a study of independent and embedded *how* and *combien* exclamative constructions as they appear in the English to French (1,410,121 words) and French to English (1,179,530 words) directional sub-corpora of Europarl Direct (Cartoni et al. 2013). These corpora were extracted from the Europarl corpus (Koehn 2005), made up of transcribed speeches given in the European Parliament, and the translation of these transcriptions. Both exclamatives are examined in terms of syntax, degree, and performativity, before the manner in which they have been translated is quantified and explored. It is important to note that in this study, this data is seen as constituting spoken language. It is possible that the transcription process allowed for some "cleaning-up" of the original speech, as pauses and redundant repetition do not appear in the data. However, while it may not be spontaneous speech, the original speeches were intended to be given orally and so likely to make effective use of the highly expressive exclamative. Furthermore, if the directional data is a question of the translation of written transcriptions, not of an oral interpretation of the original speech, it is likely to be more complete and thus more fit-for-purpose.

The question may remain as to why the French exclamative *combien* has been chosen as the point of comparison in this study. It has been highlighted that English how exclamatives may be expressed by a multitude of constructions (*combien, comme, que, ce que, qu'est-ce que*) (Jones 1996: 519), and, furthermore, *combien* has been described as "highly literary" and "archaic" and so may have seemed unlikely to occur often in the corpora at hand (Jones 1996: 519; Marandin 2018: 48). The decision to compare *how* with *combien* is based on the initial analysis of how *how* exclamatives are translated into French. In the data, *combien* was in fact chosen to express the *how* exclamative more than any other word, and so was selected to be examined in the French data.

This choice is further sustained by the preliminary results of this study. The data shows an interesting yet expected overlap in terms of the degree modification proposed by the exclamative words and the constructions they are used in, while the French shows stronger performativity. Provisional results also seem to provide an interesting contrast to those of the aforementioned study into *what* and *quel* exclamatives. More independent, clause initial *how* or *combien* exclamatives were omitted in translation than translated using an equivalent exclamative construction. Furthermore, there appears to be a slight directional difference, as more embedded *how* exclamatives were omitted when translated into French than embedded *combien* exclamatives were when translated into English.

References

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

Cartoni B., S. Zufferey & T. Meyer. 2013. Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics* 27: 23-42.

Collins, P. 2005. Exclamative clauses in English, *Word*, 56 (1), 1-17.

Grevisse, M. & A. Goosse. 2008. *Le bon usage: grammaire française.* Brussels: De Boeck University.

Heine, B., G. Kaltenböck & T. Kuteva. 2020. On the status of wh-exclamatives in English. *Functions of Language.* 272: 207-233.

Huddleston, R. 1984. *Introduction to the Grammar of English.* Cambridge: Cambridge University Press.

Huddleston, R. & G. K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Jones, M. A. 1996. *Foundations of French Syntax*. Cambridge: Cambridge University Press

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit 5: 79-86.
Marandin, J. 2010. Les exclamatives de degré en français. *Langue française.* 1 (165), 35-52.
Marandin, J. 2018. La phrase exclamative et l'exclamation en français contemporain [PDF file]. Retrieved from llf.cnrs.fr/sites/llf.cnrs.fr/files/u63/phrase_exclamative.pdf.
Michaelis, L. A. 2001. Exclamative constructions, in: M. Haspelmath, E. König, W. Oesterreicher & W. Raible (eds*.) Language Typology and Language Universals.* Berlin: de Gruyter, 1038-1050.
Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language.* London: Longman.
Rett, J. 2008. A degree account of exclamatives. In Gibson & Friedman (eds.). Proceedings of Semantics and Linguistics Theory XVIII. Ithaca, NY: CLC Publications.
Rett, J. 2011. Exclamatives, Degrees and Speech Acts. *Linguistics and Philosophy* 34. 411–442.
Riegel, M., J-C. Pellat, & R. Rioul. 2009. *Grammaire méthodique du français*. Paris: PUF.
Siemund, P. (2017). English exclamative clauses and interrogative degree modification. In: M. Napoli & M. Ravetto (eds.), *Exploring Intensification: Synchronic, diachronic and cross-linguistic perspectives*, 207–228. Amsterdam: Benjamins
Troughton, F. (Subm). *Quelle intensité!* A study of the translation of French and English exclamatives using *quel* and *whatI.*

The fate of 'pseudo-' words: a contrastive corpus-based analysis

Kristel Van Goethem[1], Muriel Norde[2], Francesca Masini[3]
F.R.S.-FNRS & Université catholique de Louvain[1], Humboldt University Berlin[2], University of Bologna[3]
kristel.vangoethem@uclouvain.be, muriel.norde@hu-berlin.de, francesca.masini@unibo.it

Evaluative morphology is by now a well-established domain of investigation (e.g. Bauer 1997, Grandi & Körtvélyessy 2015). However, while large-scale typological studies are available (e.g. Körtvélyessy 2015), in-depth contrastive studies that aim at comparing specific formatives in different languages are still scarce. In addition, evaluative morphology has essentially been restricted to the expression of diminution and augmentation, and their pejorative or ameliorative connotations (e.g. Dressler & Barbaresi 1994, Bakema & Geeraerts 2000), whereas other crucial functions such as the expression of 'approximation' (in a broad sense) have been largely ignored (Masini & Micheli 2020). The present case study, which forms part of a broader collaborative project on approximation in morphology, intends to bridge this double gap by examining the formal and semantic properties of the prefix *pseudo-* in 8 European languages: Danish, Dutch, English, German, French, Italian, Spanish and Swedish.

Pseudo- constitutes an interesting case due to its historical development. Originally a compounding element in Ancient Greek (e.g. in *pseudologia* 'a false speech'), *pseudo-* was borrowed into a variety of European languages. Initially, Greek compounds were borrowed as a whole (e.g. English *pseudonym*). Later, *pseudo-* came to be combined with native words, mostly pertaining to (spurious) science (e.g. Dutch *pseudo-deskundige* 'pseudo-expert', *pseudowetenschappelijk* 'pseudoscientific'; *WNT*). Contemporary data, however, show that *pseudo-* is expanding to collocational contexts in which the prefix is used to convey approximation, typically with an ironic tone or negative connotation (1-5).

(1) *Il fatto è che il calcio* […] *è diventato in Italiai* [sic] *uno sport per vecchi ricchi e per* ***pseudotifosi*** *in pantofole*. 'The thing is that, in Italy, football has become a sport for rich old people and pseudo-supporters in slippers' [Italian]
(2) *This must be a new trend in* ***pseudo left*** *thinking, a total failure to understand basic logic.* [English]
(3) *Een* ***pseudo historische roman*** *gebaseerd op oeroude Britse bronnen* 'a pseudo historical novel based on age-old British sources' [Dutch]
(4) *psykoanalytikeren Otto Rank, der er lidt mindre* ***pseudo*** *end Freud og Jung*. 'the psychoanalyst Otto Rank, who is a bit less pseudo than Freud and Jung.' [Danish]
(5) *Rodeamos de* ***pseudos filósofos, pseudos intelectuales, pseudos artistas***. 'We are surrounded by pseudo-philosophers, pseudo-intellectuals, pseudo-artists.' [Spanish]

At the morphological level, we observe that *pseudo-* combines with both nouns (1) and adjectives (2). Of particular interest are constructions where *pseudo-* takes scope over a noun phrase (3), or where *pseudo* is developing into an independent adjective meaning 'fake', as in (4), a process known as 'debonding' (e.g. Norde & Van Goethem 2018). Example (5) shows that debonding may even result in morphological reanalysis as an adjective, including adjectival inflection.

In order to gain a better understanding of the morphosyntactic behaviour and productivity of *pseudo-* words in contemporary European languages, we carry out a cross-linguistic analysis based on 1000-token samples per language extracted from the TenTen web corpora (Kilgarriff et al. 2014). All relevant occurrences are annotated for their morphological properties, and their productivity is calculated based on type/token ratio and potential productivity scores. In particular, we focus on (i) the construction type, (ii) the productivity of the prefix, and (iii) its degree of debonding.

These data lead to an extensive comparative analysis of the synchronic use of *pseudo-* in the 8 languages in our study, while taking into account the process of morphological adaptation specific to each of the receiving languages (cf. Seifart 2015 and Gardani 2020 on morphological borrowing).

Based on previous research into Italian and Dutch approximative morphemes (Masini & Micheli 2020, Van Goethem & Norde 2020) and into debonding in English, French and Dutch (Van Goethem & De Smet 2014), we address the following research questions:

> RQ1: Do language-specific properties such as morphological cohesion and inflection affect the degree of debonding of *pseudo-*?
> RQ2: Is there a correlation between the morphological productivity of *pseudo-* in a particular language and its degree of debonding?

More generally, we will explore whether the morphological properties of *pseudo-* confirm the grammaticalization clines from synthetic to analytic languages, as observed for the Germanic and Romance language families (Lamiroy 2011) and compare the integration of *pseudo-* in Germanic and Romance. In addition, we will perform linear regression modelling (Levshina 2015) to establish whether there is a correlation between productivity and debonding.

Preliminary analyses suggest that languages differ considerably both in terms of type/token ratios and debonding ratios, presumably because of typological differences. For instance, the Romance languages present a higher degree of debonding than the Germanic languages, which may be due to a lower degree of morphological cohesion in the former language family. In addition, the formal resemblance of *pseudo-* with Spanish and Italian adjectives ending in *-o* may facilitate debonding and adjectival reanalysis.

References

Bakema, P. & Geeraerts, D. (2000). Diminution and augmentation. In G. Booij et al. (eds.) *Morphology. An International Handbook on Inflection and Word Formation*. Berlin: de Gruyter, 1045-1052.

Bauer, L. (1997). Evaluative morphology: In search of universals. *Studies in Language* 21(3), 533-575.

Dressler, W. U. & Barbaresi, L.M. (1994). *Morphopragmatics. Diminutives and intensifiers in Italian, German, and other languages*. Berlin & New York: Mouton de Gruyter.

Gardani, F. (2020). Borrowing matter and pattern in morphology. An overview. *Morphology* 30, 263-282.

Grandi, N. & Körtvélyessy, L. (eds.). (2015). *Edinburgh Handbook of Evaluative Morphology*. Edinburgh: Edinburgh University Press.

Kilgarriff, A. et al. (2014). The Sketch Engine: ten years on. *Lexicography* 1, 7-36.

Körtvélyessy, L. (2015). *Evaluative morphology from a cross-linguistic perspective*. Cambridge: Cambridge Scholars Publishing.

Lamiroy, B. (2011). Degré de grammaticalisation à travers les langues de la même famille. *Mémoires de la Société de linguistique de Paris* 19, 167-192.

Levshina, N. (2015). How to do Linguistics with R. Data exploration and statistical analysis. Amsterdam & Philadelphia: Benjamins.

Masini, F. & Micheli, S. (2020). The morphological expression of approximation: the emerging *simil-* construction in Italian. *Word Structure* 13(3), 371-402.

Norde, M. & Van Goethem, K. (2018). Debonding and clipping of prefixoids in Germanic: Constructionalization or constructional Change? In G. Booij (ed.) *The Construction of Words*. Cham etc.: Springer, 475-518.

Seifart, F. (2015). Direct and indirect affix borrowing. *Language* 91(3), 511-532.

Van Goethem, K. & De Smet, H. 2014. How nouns turn into adjectives. The emergence of new adjectives in French, English and Dutch through debonding processes. *Languages in Contrast* 14(2), 251-277.

Van Goethem, K. & Norde, M. (2020). Extravagant "fake" morphemes in Dutch. Morphological productivity, semantic profiles and categorical flexibility. *Corpus Linguistics and Linguistic Theory* 16(3), 425-458.

WNT = Woordenboek der Nederlandse Taal: http://gtb.inl.nl

The impact of directionality on self-repairs in English<>Chinese simultaneous interpreting: A corpus-based analysis

Xiaoyi Zhai
Swansea University
joanne.xiaoyi.zhai@gmail.com

The topic of directionality has long been a contentious issue in interpreting studies (e.g. Déjean Le Féal, 2005; Denissenko, 1989; Donovan, 2005; Seleskovitch, 1968). The debate on directionality in conference interpreting can be traced back to the different standpoints held by the "Paris School" which believes that interpreting from B to A is of higher quality and the "Soviet School" which argues that interpreting into either direction is acceptable (Pöchhacker, 2016). Though a fair amount of studies have addressed the issue of directionality from both theoretical and empirical perspectives (e.g. Chang, 2005; Gile, 2005), the impact of directionality on self-repairs is a topic that is underrepresented in the current research of simultaneous interpreting (SI), especially in English-Chinese language pair. In this respect, this study aims to investigate the impact of directionality on self-repairs in English<>Chinese SI based on a corpus-based analysis.

Under Levelt's framework (1983), making a self-repair mainly goes through three phases: the monitoring of one's own speech and the detection of the trouble by the interruption of the flow of the speech; hesitation and pausing; making the appropriate repair. Given the fact that repair mechanisms in L1 and in L2 might be both quantitatively and qualitatively different (Kormos, 1999), the types of repairs examined in this study are modified based on the available repair taxonomies (e.g. Kormos, 1999; Levelt, 1983; Petite, 2005), intrinsic features of SI as well as these two languages. The types of repairs analysed in this study include different repairs, error repairs (lexical and phonetic level), appropriateness repairs, mid-articulatory repairs and repair failures.

The hypotheses of this study are as follows:

1) Directionality influences the frequency of self-repairs in SI and interpreters will repair more frequently in English to Chinese SI.
2) Directionality influences the types of self-repairs interpreters make and interpreters will have more different repairs, error repairs, appropriateness repairs and mid-articulatory repairs in English to Chinese SI but not repair failures.

To test the two hypotheses mentioned above, a study was carried out with 10 students each working in two directions, namely from English (B language) to Mandarin Chinese (A language) and vice versa. The participants in this study are all native Chinese students and their average overall IELTS score is 7.5. All of them are from three universities that offer MA in English-Chinese translation and interpreting across the UK and have received at least one-term SI training before they took part in this study. The materials used consist of one English and one Chinese speech separately delivered by a native speaker at a UN high-level meeting. Each speech is roughly about 10 mins. Accordingly, a spoken corpus was purposely built, which consists of two sub corpora, one is the English original speech with Chinese interpretation versions produced by 10 student interpreters, the other one is Chinese original speech with English interpretation versions produced by the same group of student interpreters. All the data collected in this research, including interpretations in two directions, were transcribed by software and then manually checked. Each type of self-repairs was specifically annotated for analysis. The data analysis comprises recordings and a corpus of interpreting outputs in two language directions as well as retrospective interviews, prompted by looking at the source texts and listening to the interpretations directly after each

interpretation task. Quantitative analysis was used to analyse if directionality has an impact on the frequency and types of self-repairs and qualitative analysis of interpretations and interviews was adopted to find out triggers of self-repairs related to directionality.

In terms of the findings, this study failed to prove the first hypothesis that directionality influences the frequency of self-repairs in English<>Chinese SI. A paired-samples t-test was conducted to compare the frequency of self-repairs in English to Chinese and Chinese to English directions. There was not a significant difference in the scores for English to Chinese (M=1.58, SD=0.80) and Chinese to English (M=2.02, SD=0.73) directions; t(9)=-1.441, p=0.184. These results suggest that directionality does not have an impact on the frequency of self-repairs in English<>Chinese SI. As for the second hypothesis, this study partially confirmed it by finding that interpreters made more appropriateness repairs, different repairs, error repairs in English to Chinese SI but had more mid-articulatory repairs and repair failures in Chinese to English SI. On top of this, interpreters mostly made mid-articulatory repairs among all types of self-repairs regardless of interpreting direction. Through a qualitative analysis of the self-repairs in the target language, it shows that some processing-related problems such as numbers, information density, lexical access, terminology, listening and analysis might be the reasons that cause self-repairs in both directions.

Findings in this study will provide new insights on the decision-making by interpreters in the process of making repairs. It will contribute to closing some gaps in the literature regarding the impact of directionality on self-repairs in English<>Chinese SI, and present some pedagogical implications for SI training in these two directions.

References

Chang, C. (2005) *Directionality in Chinese/English simultaneous interpreting: Impact on performance and strategy use*. PhD Thesis. University of Texas.

Déjean Le Féal, K. (2005). Can and should interpretation into a second language be taught? In R. Godijns & M. Hinderdael (eds). *Directionality in interpreting: The 'retour' or the native?* Gent: Communication and cognition, 167-194.

Denissenko, J. (1989). Communicative and interpretative linguistics. In L. Gran & J. Dodds (eds). *The theoretical and practical aspects of teaching conference interpretation*. Udine: Campanotto Editore, 155-157.

Donovan, C. (2005). Teaching simultaneous interpretation into B: A challenge for responsible interpreter training. In R. Godijns & M. Hinderdael (eds). *Directionality in interpreting: The 'retour' or the native?* Gent: Communication and Cognition, 147-165.

Gile, D. (2005). Directionality in conference interpreting: A cognitive view. In R. Godijns & M. Hinderdael (eds). *Directionality in Interpreting: The 'retour' or the native?* Ghent: Communication and Cognition, 9-26.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language learning* 49(2), 303-342.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition* 14(1), 41-104.

Petite, C. (2005). Evidence of repair mechanisms in simultaneous interpreting: A corpus-based analysis. *Interpreting* 7(1), 27-49.

Pöchhacker, F. (2016). *Introducing interpreting studies* (2nd ed.). London/New York: Routledge.

Seleskovitch, D. (1968). *L'Interprète dans les conférences internationales*. Paris: Minard Lettres Modernes.

DIGITAL POSTERS

Adjective Phraseologies in Travel Journalism in English, Italian and Polish

David Finbar Brett[1], Antonio Pinna[1], Barbara Loranc[2]
University of Sassari[1], University of Bielsko-Biala[2]
dbrett@uniss.it, dedalo@uniss.it, bloranc@ath.bielsko.p

This presentation aims to explore adjective phraseologies in Travel Journalism in three different languages: English, Italian and Polish by means of a corpus-based methodology.

The beginning of the twenty-first century has seen the constant growth of academic interest in Travel Journalism, as testified to by numerous publications within the field of Journalism studies (e.g. Fürsich and Kavoori 2001, Hanusch 2010, Hanusch and Fürsich 2014, Pirolli 2019). This may be considered as one of the many effects of the socio-economic transformations occurring in the past decades that have brought the relevance of the tourism industry in the contemporary globalized economy to the attention of both the general public and academic researchers. Nevertheless, studies of the linguistic features of Travel Journalism in specific languages are not abundant (see Brett & Pinna, 2015; Brett, 2018; Durán-Muñoz, 2019; Pierini, 2009; Pinna, 2018; Pinna & Brett, 2018) and studies comparing linguistic features across languages are practically non-existent.

The current study aims to begin to redress this situation by carrying out a study of adjective phraseologies in Travel Journalism in three markedly different European languages: English, Italian and Polish. The reason why adjectives are focused on is because they play a particular role in the language of tourism and often contribute to the formation of recurrent phraseologies (Manca, 2008). This observation can easily be extended to the language of travel journalism, a field so adjacent to that of tourism that the line distinguishing the two is often blurred.

One of the features that are compared is that of ADJ+NOUN collocations, including a study of their connectivity (Brezina et al, 2015), i.e. what are the most productive collocates in each language? Another comparative perspective focuses on ADJ+NOUN collocations involving colour: which are the most/least productive colours in each language? Can similar colour collocations be found across languages? Finally, "[negative] but [positive]" adjective pairing (e.g. *basic but comfortable*) is examined. While this feature of the language of tourism is widely reported in the literature concerning English (Dann, 1996; and Edo Marzá, 2011, 2012), it may be of interest to determine whether the pattern can be found in the other languages, and whether it has a similar function.

The analysis illustrated in this paper necessitated the compilation of three comparable corpora of travel journalism for the three languages discussed: English, Italian and Polish. An attempt was made to select articles from newspapers of a comparable standing in the three respective speech communities. The authors had already compiled a collection of articles from the 'Travel' section of the British broadsheet *The Guardian* called the Guardian Travel Corpus (GTC). This consisted of a total of 1204 articles, amounting to one million tokens. These articles appeared in the online version of the newspaper (https://www.guardian.co.uk) over a period from 2006-2011. When compiling comparable corpora in Italian and Polish, the choice fell on *La Repubblica* (https://www.repubblica.it/) and *Gazeta* (https://www.gazeta.pl/), respectively, both of which are considered to be quality publications, aimed at an educated middle-class readership. Just as *The Guardian* has a 'Travel' section, *La Repubblica* has a section entitled 'Viaggi' and *Gazeta* has one called 'Podróże'.

After the 1M-word corpora were compiled, they were tagged for Part-of-speech and lemma using TreeTagger. Collocations were extracted using tailormade perl scripts and Mutual Information was the

statistical test chosen to measure the force of attraction between the two words. The collocations were extracted on the basis of the lemma, rather than word form, as otherwise it would have been practically impossible to directly compare English with Italian, and in particular Polish, as the latter languages are far more morphologically complex.

The results concerning ADJ+NOUN collocations suggest that the degree of connectivity is similar in all three. However, as regards the actual collocates, English and Polish are most similar as their most productive adjective collocates were all general high frequency words, such as GOOD, FIRST, OTHER, HIGH, NEW for the former and DOBRY, DUŻY, INNY, WYSOKI, STARY for the latter. The most connected collocates in Italian, on the other hand, were patently connected with the subject at hand: NATURALE, STORICO, TURISTICO, CULTURALE, MEDIEVALE.

The number of colour collocations was very similar across the three languages, as there were approximately 25 types in each. There was a certain amount of variation in the colours that were lending themselves to the collocations, but in all RED, GREEN and WHITE were the most productive, while YELLOW was completely absent from English and Italian, and had only one type in Polish, ŻÓŁTY SZLAK, arguably not a prototypical example of a collocation in any case. Apart from the predictable *white wine/vino bianco/biały wino* and *red wine/vino rosso/czerwony wino*, interesting counterparts included *golden sand/złoty piasek*, and *cuore verde/zielone serce*, literally "green lung", meaning an area with vegetation in a built-up area context.

The ADJ but ADJ pattern was found in all three languages. It was more frequent in English (84) and Polish (71), but also clearly present in Italian (51). Comparison of these frequencies with those of the pattern in the three respective reference corpora showed that in each case they are statistically significant, i.e. that they are a typical feature of Travel Journalism in all three languages. By categorising both adjectives in each token, where possible, as being positive or negative in terms of evaluation, the sequence "[Negative] but [Positive]" is by far the most common in all three languages.

References:

Brett, D. (2018) "Social Network Analysis and the Analysis of Collocations in the Language of Travel Journalism." In: Baumann, Tania (ed.) *Reiseführer – Sprach- und Kulturmittlung im Tourismus / Le guide turistiche – mediazione linguistica e culturale in ambito turistico*. Bern: Peter Lang, pp.183-207.

Brett, D. & Pinna, A. (2015). Patterns, fixedness and variability: using PoS-grams to find phraseologies in the language of travel journalism. *Procedia - Social and Behavioral Sciences* 198, pp. 52 – 57.

Brezina, V; McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20:2, pp. 139-173.

Dann, G. (1996) *The Language of Tourism. A Sociolinguistic Perspective*. Wallingford: CAB International.

Durán-Muñoz, I. (2019). Adjectives and their keyness: a corpus-based analysis of tourism discourse in English. *Corpora. 14*, 3, 351–378.

Edo Marza, N. (2011). A comprehensive corpus-based study of the use of evaluative adjectives in promotional hotel websites. *Odisea, 12*, 97–123.

Edo Marzá, N. (2012). Páginas web privadas e institucionales: el uso de la adjetivación en un corpus inglés-español de promoción de destinos turísticos. In J. Sanmartín Sáez (ed.), *Discurso turístico e Internet* (pp. 81–124). Iberoamericana/Vervuert.

Fürsich, E. & Kavoori, Anandam P. (2001). Mapping a critical framework for the study of travel journalism. *International Journal of Cultural Studies* 4:2, pp.149-171.

Hanusch, F. (2010). The dimensions of travel journalism: Exploring new fields for journalism research beyond the news. *Journalism Studies* 11:1, pp. 68-82.

Hanusch, F. & Fürsich, E. (2014). On the relevance of travel journalism: An introduction. In: F. Hanusch, & E. Fürsich (Eds.), *Travel Journalism. Exploring Production, Impact and Culture* (pp.1-17). Palgrave Macmillan.

Manca, E. (2008). From phraseology to culture. Qualifying adjectives in the language of tourism. *International Journal of Corpus Linguistics, 13*,3, 368–385.

Pierini, P. (2009). Adjectives in tourism English on the web: a corpus-based study. *Círculo de Lingüística Aplicada a la Comunicación CLAC*, *40*, 93–116.

Pinna, A. (2018) “Affect in the language of travel journalism”. In: Baumann, Tania (ed.) 2018. *Reiseführer – Sprach- und Kulturmittlung im Tourismus / Le guide turistiche – mediazione linguistica e culturale in ambito turistico*. Bern: Peter Lang, pp. 151-182.

Pinna, A., Brett, D. (2018). Constance and variability: using PoS-grams to find phraseologies in the language of newspapers. In: Kopaczyk, J., Tyrkkö, J. (Eds.), Applications of Pattern-Driven Methods in Corpus Linguistics. John Benjamins, Amsterdam, pp. 107--130.

Pirolli, B. (2019) *Travel Journalism. Informing Tourists in the Digital Age*. London: Routledge.

Testing the hypothesis of "translation as a catalyst for plain legislation" on the syntactic level: A comparison of different varieties of legislative Italian

Paolo Canavese[1], Laura Mori[2]
Université de Genève / TRANSIUS Centre[1], Università degli Studi Internazionali di Roma (UNINT)[2]
paolo.canavese@unige.ch, laura.mori@unint.eu

Several external variables, directly or indirectly related to plurilingual legislative drafting, may impact on the linguistic quality of legislation. They range from institutional multilingualism, legal drafting traditions, language policies, drafting guidelines, training programs to translation as an operative tool (Mori, forthcoming). The hypothesis that translation can contribute to improving the level of plainness of legislative texts has been expressed a number of times in the literature (see e.g. Schnyder 2001; Flückiger 2005; Egger and Ferrari 2016). As far as legislative Italian is concerned, it has recently received first empirical validations. Some corpus-based studies set out, for instance, to unveil the level of lexical readability of different varieties of legislative Italian, showing a higher level of accessibility of Swiss and EU compared to Italian legislative texts (Felici and Mori 2019; Canavese forthcoming). The findings of Mori (2019a) on syntax and readability go in the same direction. To date, however, no study is available on syntactic complexity of Swiss legislation.

Our study builds on these previous findings to check whether and to what extent the Italian versions of Swiss federal acts and EU directives resort to plainer syntactic choices compared to Italian domestic legislation (Italian origin) and implementing laws (EU-derived). A particular focus is placed on the comparison of the two translation-mediated contexts, which show both similarities (multilingual lawmaking process) and dissimilarities (direct vs. indirect applicability). We will also discuss how these contextual elements have an impact on the level of syntactic complexity and thus readability.

To carry out this study, we employed the three Italian corpora of the Eurolect Observatory Project (Mori 2018; 2019a) and LEX.CH.IT, a corpus of Swiss federal acts (Canavese 2019). Thanks to an NLP-based monitoring carried out through the automatic annotation tools developed by the ItalianNLP Lab at the ILC in Pisa (Dell'Orletta et al. 2013; Montemagni 2013), we were able to analyse from a quantitative perspective a number of aspects that can be considered as proxies of syntactic complexity. These aspects were divided in six groups: sentence complexity on the shallow level, nominal style, marked traits of verb morphology, subordination vs. coordination, syntactic tree complexity and aspects pertaining the information structure. To identify relevant trends, we also resorted to descriptive and inferential statistics and interpreted these data qualitatively.

The results confirm the existence of a correlation between the variables "translation" and "plain language", which is in line with the starting hypothesis. Indeed, compared to domestic and EU-derived Italian legislation, Swiss and EU legislative acts feature shorter sentences, a preference for the verbal style, a less complex syntactic tree and tend to stick to the SVO constituent order. Finally, the contrastive analysis of Swiss and EU legislative texts showed that the former have an only slightly but statistically significant lower level of syntactic complexity compared to the latter.

Comparing and contrasting different varieties of legislative languages in light of their drafting contexts and cultures makes it possible to refine our understanding of hybridization and harmonization dynamics (Mori 2019b). Ultimately, it also contributes to promoting the adoption of a plainer language in legislative texts.

References

Canavese, P. (2019). LEX.CH.IT: A Corpus for Micro-Diachronic Linguistic Investigations of Swiss Normative Acts in Italian. *Comparative Legilinguistics* 40, 44-65. doi: 10.14746/cl.2019.40.3.

Canavese, P. (forthcoming). Plain Legal Language through Translation: A Comparison of Swiss, EU and Italian Legislative Texts. In D. Leisser & L. Green (eds.) *Contemporary Approaches to Legal Linguistics*. Berlin: Lit Verlag.

Dell'Orletta, F., Montemagni, S. & Venturi, G. (2013). Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, 7-11 September, Hissar, Bulgaria*, 189-197.

Egger, J.-L. & Ferrari A. (2016). L'italiano federale svizzero: elementi per una ricognizione. *Studi italiani di linguistica teorica e applicata* 45(3), 499-523.

Felici, A. & Mori L. (2019). Corpora di italiano legislativo a confronto: dall'Unione europea alla Cancelleria svizzera. In B. Moretti, A. Kunz, S. Natale & E. Krakenberger (eds.) *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana (Berna, 6-8 settembre 2018)*. Milan: Officinaventuno, 287-304.

Flückiger, A. (2005). Le multilinguisme de l'Union européenne : un défi pour la qualité de la législation. In J.-C. Gémar & N. Kasirer (eds.) *Jurilinguistique : entre langues et droits = Jurilinguistics : between Law and Language*. Bruxelles: Bruylant, 340-360.

Montemagni, S. (2013). Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata*, 42(1), 145-172.

Mori, L. (2018). Observing Eurolects: The Case of Italian. In L. Mori (ed.) *Observing Eurolects. Corpus Analysis of Linguistic Variation*. Amsterdam: Benjamins Publishing House, 192-242.

Mori, L. (2019a). Complessità sintattica e leggibilità. Un monitoraggio linguistico per la valutazione dell'accessibilità dei testi legislativi europei e italiani. *Studi Italiani di Linguistica Teorica e Applicata* 48(3), 627-657.

Mori, L. (2019b). Dall'armonizzazione all'ibridazione nei testi legislativi: evidenze linguistiche e manifestazioni interculturali nell'italiano tradotto. *Entreculturas* 10, 377-392.

Mori, L. (forthcoming). Prospettive di miglioramento dell'accessibilità linguistica in italiano: verso una cultura della comunicazione istituzionale citizen-centred. In F. Orletti (ed.) *Comunicare il patrimonio culturale: accessibilità comunicativa, tecnologie, sostenibilità*, "Lingua, traduzione, didattica". Milan: FrancoAngeli.

Schnyder, B. (2001). Zur Mehrsprachigkeit der schweizerischen Gesetzgebung im Allgemeinen. *LeGes* 12(3), 33-48.

Towards a diachronic turn in corpus-based contrastive linguistics. How can historical linguistics contribute?

Evie Coussé
University of Gothenburg
evie.cousse@gu.se

This paper takes up one of the challenges put forward in Hasselgård (2020: 201) for future research in corpus-based contrastive linguistics:

> Finally, now that corpus-based contrastive studies have been with us for more than 25 years, it may be time to take a diachronic perspective.

Hasselgård refers to the suggestion of Ebeling (2016) that the compilation of "parallel corpora matching the existing ones in terms of content and structure, but comprising texts of a more recent date […] would pave the way for a field of diachronic corpus-based contrastive studies". This paper ties in with this suggestion, but proposes to take it one step further, inspired by recent corpus projects in historical linguistics.

Updating an existing parallel corpus with a matching parallel corpus of more recent date gives rise to a pair of parallel corpora from different time periods that are comparable in their design. In a diachronic perspective, we are essentially dealing with a comparable corpus, with no relation of equivalence between source texts and translations across time periods. This paper suggests that an alternative corpus design allows to create such a diachronic relation of equivalence. Instead of building an entirely new parallel corpus of more recent date matching an existing parallel corpus, one adds more recent translations of the same source texts to the existing parallel corpus. Figure 1 represents a minimalistic design of such a diachronic multilingual parallel corpus.

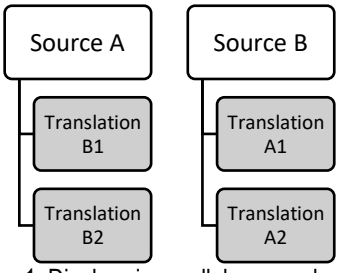


Figure 1. Diachronic parallel corpus design

The source texts in language A and B (on the top of the figure) together with their translations B1 and A1 (in the middle) represent a traditional bidirectional multilingual parallel corpus. What makes this corpus design a diachronic parallel corpus is the inclusion of translations B2 and A2 (at the bottom of the figure). They are translations from the same source text as translations B1 and A1 but from a more recent date. The translations from both time periods have a relation of equivalence with the same source text which offers a *tertium comparationis* for the diachronic comparison of the two translations.

This corpus design has been implemented in a number of recent corpus projects in historical linguistics (Chiarcos et al. 2014, Kalouli et al. 2019, Breder Birkenes et al. 2020, Bouma et al. 2020). An example is the EDGeS Diachronic Bible Corpus (Bouma et al. 2020) bringing together thirty-six Bible translations in English, Dutch, German and Swedish from the fourteenth century until the present day. All of the cited diachronic parallel corpora make use of historical Bible translations in order to cover comprehensive parts

of the written history of Germanic and Romance languages. This might not be the type of diachronic parallel corpus that is best suited for a diachronic turn in corpus-based contrastive linguistics – both in terms of desired time depth and text genre. Yet, the budding field of diachronic contrastive linguistics could benefit from a close collaboration with these existing corpus initiatives in historical linguistics to exchange experience in both compiling and analyzing diachronic parallel corpora.

References

Bouma, Gerlof, Evie Coussé, Trude Dijkstra & Nicoline van der Sijs (2020). The EDGeS Diachronic Bible Corpus. *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 5232-5239.

Breder Birkenes, Magnus, Jürg Fleischer & Stephanie Leser-Cronau (2020). A diachronic and areal typology of agreement in Germanic. *STUF* 72, 61-114.

Chiarcos, Christian, Maria Sukhareva, Roland Mittmann, Timothy Price, Jens Chobotsk & Gaye Detmold (2014). New Technologies for Old Germanic. Resources and Research on Parallel Bibles in Older Continental Western Germanic. *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 22-31,

Ebeling, Signe (2016). Does corpus size matter? Revisiting ENPC case studies with an extended version of the corpus. *Nordic Journal of English Studies* 15, 33-54.

Hasselgård, Hilde (2020). Corpus-based contrastive studies. Beginnings, developments and directions. *Languages in Contrast* 20, 184-208.

Kalouli, Aikaterini-Lida, Rebecca Kehlbeck, Rita Sevastjanova, Katharina Kaiser, Georg A. Kaiser & Miriam Butt (2019). ParHistVis: A Visualization of Parallel Multilingual Historical Data. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 109-114.

Conjunctive markers of contrast in English and French: syntactic patterns and discourse effects

Maïté Dupont
Université catholique de Louvain
maite.dupont@uclouvain.be

Conjunctive markers (CMs) are linguistic items that express logical relationships between textual units, such as *however, although* or *and* (see e.g. Martin 1992). They are usually identified on the basis of syntactic criteria. For example, one necessary condition is that they must link clauses or larger units. Elements that occur below the clause are not considered to perform a conjunctive function (e.g. Hoek et al. 2017). In addition, the textual segments related by CMs may be of various structural types (i.e. include a verb in the finite or non-finite form, or even no verb at all), or function at different syntactic ranks (i.e. main, hypotactic, embedded or minor clause, where minor clauses refer to verbless, non-finite or hypotactic clauses standing alone). However, research on CMs has customarily focused on the semantic features of these units, with very little attention paid to their syntactic patterning. The objective of this study is to shift the focus to syntax by investigating and comparing the types of syntactic segments in which English and French CMs are included, focusing on the category of contrast.

The study is based on a comparable corpus of newspaper editorials (c. 2 million words per language). Following automatic extraction of all potential English and French CMs of contrast from the corpus, the data was disambiguated manually in context to weed out irrelevant instances. The disambiguated data set (15,364 markers in total) was then coded manually for (i) type of CM (coordinator, subordinator or conjunctive adjunct); (ii) structural type of the host clause (finite, non-finite or verbless); and (iii) rank status of the host clause (main, hypotactic, embedded or minor).

The study shows that the dominant syntactic patterns of use of CMs of contrast are similar in English and French. However, it also highlights a number of significant differences between the languages. For instance, CMs used in minor clauses, as in (1), are significantly more frequent in French than in English, which displays a significantly higher proportion of subordinators used in verbless clauses than French, as in (2).

(1) Nous nous enflammons pour ou contre Dieudonné, pour ou contre le voile islamique, demain pour ou contre les coups de canif à la loi de 1905. **Mais** sur le meurtre de Van Gogh, qui nous parle exactement de la même chose, pas d'émotion, pas d'indignation, pas d'effroi.
[***But*** *on Van Gogh's murder (…) no emotion, no indignation, no terror*].
(2) Efforts at political reform in Saudi Arabia, Egypt, Jordan and Kuwait, **while** welcome, are but partial remedies.

A qualitative analysis of the results reveals that the patterns identified are not equivalent regarding the discourse effects that they produce. For example, the use of a coordinator or conjunctive adjunct in a minor clause – as in (1) – typically lays emphasis on both the relation of contrast expressed by the marker, and the segment introduced by it. Such patterns of use contribute to increasing the 'punchy' character of a text, in line with the highly persuasive tone of the editorial register (e.g. Biber 1988: 148). In other words, syntactic choices related to the use of conjunctive markers reflect broader (differences in) strategies of textual development in English and French, which highlights their role at the syntax-discourse interface.

References

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
Chuquet, H. & Paillard, M. 1987. *Approche linguistique des problèmes de traduction anglais-français*. Paris: Ophrys.
Hoek, J., Evers-Vermeul, J. & Sanders, T. 2018. Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory* 14(2): 357–386.
Martin, J. 1992. *English Text: System and Structure.* Amsterdam: John Benjamins.

CHEU-LEX: a parallel multilingual corpus of Swiss and EU legislation

Annarita Felici[1], Antonio Giovanni Contarino[2], Francesco Fernicola[2], Adriano Ferraresi[2], Silvia Mattiuzzi[1], Silvia Polito[2]
Faculty of Translation and Interpreting - University of Geneva[1], Department of Interpreting and Translation- University of Bologna[2]
annarita.felici@unige.ch, antonio.contarino@studio.unibo.it, francesco.fernicola@studio.unibo.it, adriano.ferraresi@unibo.it, Silvia.Mattiuzzi@etu.unige.ch, silvia.polito2@studio.unibo.it

In this poster, we present the design and compilation of the CHEU-LEX corpus, a parallel and comparable corpus of Swiss and EU legislation in the three official languages of the Confederation (French, German and Italian). Although Switzerland is not part of the EU, it has close relations with it, and is indirectly affected by its decisions via bilateral agreements since 1972. The corpus was built for two main aims. On the one hand, it was conceived to investigate the influence of EU drafting and translation practices on Swiss legislation, following in the wake – and expanding the scope – of similar initiatives focusing on legislation of EU countries (e.g., Biel 2014; Mori 2018). On the other hand, it aims at providing a richly annotated multilingual resource to explore legislative language at several levels (macro-textual, lexical, morphosyntactic) and according to different perspectives (monolingual, cross-lingual and cross-textual).

CHEU-LEX consists of bilateral agreements between Switzerland and the EU in the time span 1972-2017 and of Federal legislation representing the reception of these agreements, for a total of 792 texts and approximately 4.2 million tokens. The corpus is made publicly available through the NoSketchEngine platform[12] and involves several levels of annotation focusing on: a) contextual information on date of publication and topic; b) structural features (title, preamble, articles, annexes), c) Parts-of-Speech (POS); d) syntactic dependencies; and e) sentence alignment.

The poster will provide details on, and discuss issues that arose at, each phase of corpus construction. Specifically, after downloading legislation from the Swiss Federal Law website, texts were annotated with structural and contextual information by means of a Perl script and then segmented and aligned at sentence level using Intertext Editor (Vondřička 2014). Given the structure of legislative texts and typical sentence arrangement, the output of the segmentation software had to be corrected manually. A lower level of segment granularity was applied to the titles and subheadings of legal texts, where in absence of punctuation, we merged the various segments for the sake of contextual meaning. In the next phase, POS tagging and lemmatisation were carried out using the Italian and French TreeTagger (Baroni 2007; Schmid 1994), and the German RFTagger (Schmid 2008). The raw output was manually revised and corrected, and elements were added to the tagsets to account for textual features typical of legal texts, including abbreviations, foreign words, list markers and abrogated elements. Dependency Parsing was carried out using SpaCy (Honnibal 2020), with Italian and French tags based on Universal Dependencies (UD Italian ISDT v2.5 and UD French Sequoia v2.5), and German ones based on the TIGER Corpus. The output of the SpaCy parser did not undergo any manual revision.

To illustrate the potential of the corpus to investigate typical features of multilingual and translated legislative texts, the poster will showcase queries exploiting all levels of annotation in CHEU-Lex, according to three perspectives: monolingual (bilateral agreements in a single language), cross-lingual (bilateral agreements in the three languages) and cross-textual (bilateral agreements and Swiss legislation in the same language).

[12] http://corpora.fti.unige.ch/crystal/

References

Baroni M., Schmid H., Zanchetta E., Stein A. (2007): "The Enriched TreeTagger System", in *Proceedings of the Evalita Workshop (10th Congress of Italian Association for Artificial Intelligence, AI*IA 2007),* University of Roma "Tor Vergata", Rome.

Biel, L. (2014): *Lost in the Eurofog: The Textual Fit of Translated Law,* Frankfurt, Peter Lang.

Honnibal M., Montani I., Van Landeghem S., Boyd A. (2020): "spaCy: Industrial-strength Natural Language Processing in Python", Zenodo, url:https://doi.org/10.5281/zenodo.1212303.

Mori, L. (2018): *Observing Eurolects. Corpus analysis of linguistic variation in EU law,* Amsterdam, John Benjamins.

Schmid H. (1994): "Probabilistic Part-of-Speech Tagging Using Decision Trees", in *Proceedings of International Conference on New Methods in Language Processing*, Manchester. 44-49.

Schmid H., Laws F. (2008): "Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging", in *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester. 777-784.

Vondřička P. (2014): "Aligning parallel texts with InterText", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA). 1875-1879.

How did Artusi's *La Scienza in Cucina e l'Arte di Mangiar Bene* make it across the Atlantic: translation and adaptations into Brazilian Portuguese

Anabela Cristina Costa da Silva Ferreira[1], Stella E. O. Tagnin[2]
University of Bologna[1], University of São Paulo[2]
anabela.ferreira@unibo.it, seotagni@usp.br

There is a certain belief that European Portuguese and Brazilian Portuguese are not very different. This issue has already been broached by Tagnin and Teixeira (2004) who showed that there are striking terminological differences in the culinary domain. Their findings were based on a corpus of recipes. The aim of this paper is to take this analysis a step further by comparing two Portuguese translations of Pellegrino Artusi´s *La Scienza in Cucina e l'Arte di Mangiar Bene* (1891). The first translation, not published (hereafter PT-ANA), is by Anabela C. C. da Silva Ferreira (Artusi, s/d), a translator of Portuguese origin living in Italy, and was the basis for the Portuguese translation published in Brazil in 2009 (Artusi) (hereafter PT-PUB). These two versions plus the original Italian, were digitalized and constitute our parallel aligned corpus analyzed with WordSmith Tools 7 (Scott, 2006). To pursue a broader analysis, we will address the following questions: 1. Which features underwent the most adaptations in the Brazilian published translation as compared to the unpublished translation? 2. How can one infer cultural aspects based on translation strategies?

The paper begins with an Introduction presenting a short overall comparison between the three versions, discussing especially the introductory texts added to the published Portuguese translation and their relevance to how the text was rendered to the Brazilian audience. Section 2 presents a comparative analysis based on the KeyWords of each version. For the purpose of this paper the analysis focuses on the first 107 recipes, which cover the sections BRODI, GELATINA E SUGHI (recipes 1-6) and MINESTRE (recipes 7-107). The keywords were extracted by comparing the wordlists of each corpus with the wordlists of corresponding general language reference corpora. A qualitative analysis comparing PT-PUB and PT-ANA keywords highlighted significant terminological differences, which were looked up in the aligned corpora. Each translation was compared with the Italian original to identify translator strategies, and the two Portuguese translations were compared to each other to detect changes made in the Brazilian published one. The results are discussed in Section 3 and revealed that the differences between the two Portuguese translations are not only terminological, but also lexicogrammatical and stylistic, in addition to spelling distinctions.

The Brazilian translation is preceded by five texts in which authorities related to the Emilia-Romagna region attest to the importance of the project. These introductions seem to be clear evidence of the relevance given to this publication as well as the Brazilian formality that surrounds this enterprise. Ferreira`s main objective was to make available to the large contingent of Italian immigrants in Brazil "a publication that speaks of roots, traditions and family" (Ferreira, 2017, p. 181). To that end the translator claims to have tried to be as faithful as possible to the Italian work so that it could arrive at each Brazilian family of Italian descent, creating a solid, familial and historic bridge with a common desire of gastronomic conviviality (Ferreira, 2017, p. 184). This included recreating, in the Portuguese language, Artusi`s writing style. However, because this would sound quite peculiar to today`s reader, the Brazilian editor opted to simplify the language, making it more accessible to the target audience. Although Ferreira is a native of Portugal, she made an effort to produce a text in Brazilian Portuguese. Nevertheless, her native variant shone through in certain syntactic constructions and culinary terms, which also demanded a revision of the text to adapt it to Brazilian conventions.

To make reading easier for that audience, footnotes were added referring, for example, to Italian towns, literary works and types of pasta. Although Artusi features a glossary of regional Emilian-Romagnoli terms which was also translated, an additional glossary was included by the Brazilian editor to explain certain cooking procedures no longer in use, typical ingredients and a variety of types of pasta. Another feature denoting the editor's desire to make the target audience actually try out the recipes is the layout of these pages which feature a blank space with lines on the outer edge of the recipes for annotations by the reader.

To exemplify our findings, we will show the results based solely on the analysis of recipe n. 46 (*Cuscussù*), which have shown an array of alterations, confirming that the two variants of Portuguese present significant differences in the gender of recipes (Tagnin & Teixeira, 2004), but also in other aspects. Here are some lexical differences:

PT-ANA	PT-PUB	EN
Aipo	Salsão	celery
Taça	Tigela	dish
torteira de cobre	Forma de cobre	copper baking pan
prato de sopa	Prato fundo	soup bowl
Fumo	Vapor	steam
Legumes	Verduras	vegetables
Dose	Quantidade	amount

The Keyword list of PT-ANA highlighted verbs in the plural, such *as coloquem* and *piquem*. These were replaced by their singular form, as is usual in Brazilian recipes. One verb in particular, DEITAR, is solely used in European Portuguese in this context, and was replaced by COLOCAR or JOGAR, depending on the context. *Couve lombarda* (*cavolo verzotto* in Artusi's text) is unknown in Brazil and became simply *repolho* (cabbage). Syntax was also adapted:

PT-ANA	PT-PUB	EN
uma hora e quarto	uma hora e quinze	an hour and fifteen minutes
cozinhar ao vapor	cozinhar no vapor	cook over steam
a meio da cozedura	no meio do cozimento	halfway through the cooking process
a refogar	para refogar	by sauteing

The Brazilian translation also adapted the 'list of ingredients' to our conventional format, that is, the amount preceding the ingredients: *750 gramas de peito de vitela* as opposed to *Um pedaçinho de peito de vitela, 750 gr.* in PT-ANA, which emulates Artusi`s *Spicchio di petto di vitella, grammi 750*.

However, a few slips, like *espinafres* for *espinafre* and *conserva de tomate* instead of *extrato de tomate*, remained.

At this point one could say that Ferreira`s translation was source-oriented, or foreignizing in Venuti`s (1998) terms, while the Brazilian version is clearly target-oriented, or domesticating. As regards the differences between the two Portuguese variants, our analysis has shown that they are significant and require near-native knowledge of the target culture to ensure an acceptable translation.

References

Artusi, P. (1891). *La Scienza in Cucina e l'Arte di Mangiar Bene.* Firenze: Salvadore Landi.

Artusi, P. (s/d). *A ciência na cozinha e a arte de comer bem - manual prático para as famílias*. (A. C. Ferreira, Trad.) Manuscript.
Artusi, P. (2009). *A Ciência na Cozinha e a Arte de Comer Bem* (1a. edição). (A. C. Ferreira, & M. O. Bertolozzi, Trads.), Itu, São Paulo, Brasil: Associação Emiliano-Romagnola Bandeirante.
Ferreira, A. C. (2017). "A ciência na cozinha e a arte de comer bem" de Pellegrino Artusi - Um símbolo da unificação linguisti no Brasil. *Dialogoi - rivista di studi comparativici*, 171-187.
Scott, M. (2006). *WordSmith Tools*. Acesso em 9 de junho de 2016, disponível em http://www.lexically.net/wordsmith/downloads/
Tagnin, S. E., & Teixeira, E. D. (2004). British vs. American English, Brazilian vs. European Portuguese - how close or how far apart? A corpus-driven study. *Lodz Studies in Language vol. 9 - Practical Applications in Language and Computers - PALC 2003* (pp. 193-208). Frankfurt am Main: Peter Lang.
Venuti, L. (1998). *The Scandals of Translation: Toward an Ethics of Difference.* London & New York: Routledge.

A corpus-based comparison of prosodic features in on/off-screen dubbing

Erick García Chávez[1], Alina Karakanta[1,2]
Università di Trento[1], Fondazione Bruno Kessler[2]
erickgch@gmail.com, akarakanta@fbk.com

The importance of prosody in dubbing has been widely studied from a naturalness perspective, often in comparison to spontaneous speech (Baños-Piñero and Chaume, 2009; Romero-Fresco, 2006; Sánchez-Mompeán, 2020). Dubbing is a highly constrained type of translation (Chaume, 2020) and the presence of an actor's face on screen pushes synchronisation constraints to their extreme. For instance, Pettorino and Vitagliano (2003) found that dubbers increased the duration of silence in close-up shots to achieve synchrony between the shorter English translation and the original Italian version. However, no previous work has studied whether the visibility of the actor's lips influences the prosody of the rendered dubbed speech. We investigate whether on-screen speech constrains dubbing actors to imitate the prosody of the source text, by examining whether prosodic elements between source and target are more similar for on-screen speech.

We conduct a corpus-based study on Heroes (Öktem, 2018), a phonetically and prosodically annotated corpus of English-to-Spanish dubbed speech, containing 7000 utterances - to our knowledge, the only parallel dubbing corpus offering such annotations. We use the annotation of Karakanta et al. (2020) to separate the segments depending on whether the actor's face is visible (ON), completely non-visible (OFF) or visible only for a part of the utterance (MIXED). We look at English-Spanish differences in 1) variation of mean fundamental frequency (f0; in cents) and intensity (in dB) per segment, 2) concurrent frequency and intonation peaks in paired segments and 3) prosodic phrase length between source and target in terms of duration (in seconds) and number of phrases per segment.

While a greater mean f0 variation is observed for Spanish, and English shows a higher intensity variation, these differences are nonetheless not significant. Moreover, we find that concurrent frequency peaks are more frequent in OFF and MIXED dialogue, while for intensity, they are more frequent in ON speech. This contradicts our expectations, i.e., that ON would contain the highest percentages in both cases. Looking at prosodic phrase length, we observe a negligible mean difference between ON and OFF, while MIXED dialogue shows relatively larger values. This suggests that, in this corpus, length constraints on the dubbed speech do not vary depending on whether the dialogue is on-screen or not.

We found that screen category does not affect voice modulation in dubbing in Heroes corpus; the differences between languages are possibly due to differences in the prosody of English compared to Spanish and among individual speakers. Additionally, we observed that regardless of mode, frequency/intensity peaks are replicated in less than 10% of sentences. Our findings are in line with Sánchez Mompeán (2020), who argues that despite some patterns introduced by ST assimilation, dubbing actors resort to the Spanish prosodic repertoire. Finally, voice modulation may not be possible when translation/synchronization strategies do not vary depending on screen category, as corroborated by Karakanta et al. (2020) who did not observe distinctive features between ON/OFF at the textual level. Further research will investigate in-depth the role of translation and synchronisation strategies on voice modulation in dubbing.

References

Baños-Piñero, R & Chaume, F. (2009). Prefabricated orality a challenge in audiovisual translation. In: G. Nadiani M. Giorgio Marrano and C. Rundle, editors, *inTRAlinea Special Issue: The Translation of Dialects in Multimedia*. inTRAlinea.

Chaume, F. (2020). Dubbing. In: *The Palgrave handbook of audiovisual translation and media accessibility*. Palgrave Macmillan, Cham, pp. 103–132

Karakanta, A., Bhattacharya, S., Baumann, T., Nayak, S., Negri, M. & Turchi, M. (2020). The Two Shades of Dubbing in Neural Machine Translation. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4327–4333.

Öktem, Alp. Heroes Corpus. 2018 http://hdl.handle.net/10230/35572

Pettorino, M., & Vitagliano, I. (2003). Prosodic characteristics of dubbed speech. In: *Proceedings of the XV International Congress of Phonetic Sciences*, Barcelona, Spain, pp. 2865–2868.

Romero-Fresco, P. (2006). The Spanish Dubbese: A case of (un)idiomatic friends. *JosTrans*, 6, 134-151.

Sánchez Mompeán, S. (2020). *The Prosody of Dubbed Speech - Beyond the Character's Words*. Palgrave Studies in Translating and Interpreting. https://doi.org/10.1007/978-3-030-35521-0

A contrastive study of EN *such* and FR *tel*

Lobke Ghesquiere
University of Mons
lobke.ghesquiere@umons.ac.be

English *such* and the corresponding French *tel* are very versatile, both semantico-pragmatically and syntactically, being able to express, for instance, both textual, cohesive relations and expressive, emphatic meanings, as in the (a) and (b) examples respectively.

(1) a. In fact CIA and KGB officers entertain each other frequently in their homes. </s><s> The CIA's files are full of mention of **such** relationships in almost every African station. (enTenTen15)
b. The sequence was one of those Easter hymns in which Luther took **such** delight. (enTenTen15)

(2) a. *énormément de forums exigent l'acceptation d'une charte avant toute participation. … Une* ***telle*** *charte régit l'usage qui est fait du forum de discussion* (frTenTen17)
'very many fora require the acceptance of a charter before any participation … Such a charter registers the use that is made of the discussion forum'
b. *Bilel* … ne s'attendait pas à une* ***telle*** *méconnaissance de la loi parmi les élèves dont il a la responsabilité.* (frTenTen17)
'Bilel* … hadn't expected such ignorance of the law among the students he was responsible for.'

It is not surprising then perhaps that both items have received considerable attention in the literature. For English *such*, landmark studies include a.o. Bolinger (1972), Altenberg (1994) and Spinillo (2003). French *tel* has been studied and described in a.o. Van Peteghem (1995) and Riegel (1997). To the best of our knowledge, however, no contrastive or translation-based studies have been carried out on the two items. This study aims to do just that. We will look into the range of prenominal uses of *such* and *tel*, charting their specific functional profiles and assessing their translational equivalence.

On the basis of monolingual data, the descriptions of *such* and *tel* available in the literature will be verified, paying particular attention to the structural and collocational distributional patterns of their textual and expressive uses. For instance, do the gradable nouns used with expressive *such* and *tel* fall into specific collocational sets? Do the different functions of *such* and *tel* correlate with different modifiers in the noun phrase?

On the basis of translation data, taken from the parallel English and French EUROPARL corpora, we will assess the intertranslatability of *such* and *tel*. The bilingual Collins, Cambridge and Larousse dictionaries all list *such* as the preferred translation for *tel* in all its uses, and vice versa. The only exception is the translation of *such* modifying an adjective, where French requires the use of *si*, as in (3).

(3) a. This is such an important matter for consideration that firm conclusions have to drawn. [sic] (EUROPARL7, English)
b. *C'est un aspect si important qu'il faut le reporter de l'exposé des motifs aux conclusions*. (EUROPARL7, French)

Preliminary data study, however, has shown that translators are far more creative when translating *such* and *tel* into French and English respectively. In the English-to-French translation data, for instance, prenominal emphatic *such* is translated by *tel*, but more frequent translations include *autant* and *aussi*.

Translations of textual, phoric *such*, by contrast, often do not feature *tel* but rather a form of the demonstrative *ce* or the article *le*. This study will inventory the different translations and try to uncover specific translation strategies, paying particular attention to possible factors influencing translation choices such as emphatic strength, type of phoric relation, structural realization and collocation.

Primary sources

Cambridge dictionaries. Available online at https://www.dictionary.cambridge.org Collins dictionaries. Available online at https://www.collinsdictionary.com.
English Web 2105 (enTenTen15) corpus. Accessed through SketchEngine, available online at https://www.sketchengine.eu/
EUROPARL corpora. Accessed through SketchEngine, available online at https://www.sketchengine.eu/
French Web 2017 (frTenTen17) corpus. Accessed through SketchEngine, available online at https://www.sketchengine.eu/
Larousse dictionaries. Available online at https://www.larousse.fr.

Secondary sources

Altenberg, Bengt. 1994. On the functions of *such* in spoken and written English. In Nelleke Oostdijk, Pieter De Haan & Jan Aarts (eds.). Corpus-based research into language. Amsterdam: Rodopi. 223–239.
Bolinger, Dwight. 1972. *Degree Words*. The Hague: Mouton.
Riegel, Martin. 1997. Tel adjectif: Grammaire d'une variable de caractérisation. *Langue française* 16: 81–99.
Spinillo, Mariangela Galvão. 2003. On such. *English Language and Linguistics* 7: 195–210.
Van Peteghem, Marleen. 1995. Sur les emplois anaphoriques de *tel*. *Sémiotiques* 8: 57–78.

Interpreted discourse or the discourse of interpreters? A corpus-based investigation of interpreters' individual language use

Andrea Götz
Károli Gáspár University
drgoetz.a@gmail.com

In recent years, the use of connectives (e.g. *so, however, but*) has been examined in corpus-based interpreting studies from several angles. It has been found, contrary to prior expectations, that the frequency of connectives increases in interpreting and serves the re-structuring of interpreted discourse (Defrancq et al. 2015). Furthermore, the frequency of connectives in interpreting can also exceed that of original, comparable discourse (Defrancq 2018), while some of their functions may also be over-represented in interpreting (Defrancq 2016). Male interpreters appear to deploy these items more frequently than female interpreters, although this difference is not statistically significant, on the other hand, the delivery speed of speaker correlates significantly with the frequency of connective in interpreted discourse (Magnifico and Defrancq 2020). Due to the central role connectives play in re-creating cohesion in interpreted speech, it is vital to understand their behaviour in interpreted discourse. However, the aforementioned results have not yet been widely corroborated in other language pairs, and do not account for potential individual variation among interpreters. This descriptive study addresses this gap by examining connective use in the discourse output of seven individual interpreters, working from Hungarian into English in different modes of interpreting (direct native, retour, relay) at the European Parliament (EP).

While the question of individual variation and preferences of interpreters have been studied with regard to various variables, these studies concerned the output of trainee interpreters (Baxter 2019, Gumul 2006), and/or microcorpora consisting of the interpreting output of two male interpreters in samples of a small number of speeches (Kajzer-Wietrzny 2012) without providing a systematic description of connective use (Van Besien and Meuleman 2008). As a result, the influence of individual variation among interpreters on connective use is still not satisfactorily explored.

The corpus of this study contains the interpreting output of two female Hungarian to English direct interpreters (26 min 39 sec, 5151 words, 12 speeches; 28 min 47 sec, 4535 words, 13 speeches), three male Hungarian to English retour interpreters (21 min 40 sec, 2737 words, 10 speeches; 19 min 26 sec, 2675 words, 10 speeches; 19 min 59 sec, 2647 words, 10 speeches), and two male relay interpreters (interpreting from English to German to Hungarian) (16 min 24 sec, 1531 words, 8 speeches; 11 min 1 sec, 1182 words, 5 speeches) sourced from the Hungarian European Parliamentary Interpreting Corpus (HEPIC). The corpus of this study thus comprises altogether 2 hours 33 minutes, or 20,458 words. The study examines the use and frequency of a set of connectives *(as a result, but, however, nevertheless, now, so, that is, why, therefore, though, thus, yet)* with regard to individual variation.

According to the results, significant differences persist between the connective use of individual interpreters. While connective frequency varies greatly among interpreters, all interpreters except two (retour interpreters) add more items than they translate. Due to the limited size of this study, it offers only preliminary findings on the topic, nevertheless, it raises questions about the necessity of accounting for individual variation.

References

Baxter, R.N., 2019. Style versus strategy in simultaneous interpreting: Different approaches and their effects. *Quaderns* 26, 287–305.

Defrancq, B., 2018. The European Parliament as a discourse community: its role in comparable analyses of data drawn from parallel interpreting corpora. *The Interpreters' Newsletter* 115–132. https://doi.org/10.13137/2421-714x/22401

Defrancq, B., 2016. Well, interpreters… a corpus-based study of a pragmatic particle used by simultaneous interpreters, in: Corpas Pastor, G., Seghiri, M. (Eds.), *Corpus-Based Approaches to Translation and Interpreting*. Peter Lang, Bern, pp. 105–128. https://doi.org/10.3726/b10354

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

Gumul, E., 2006. Explicitation in Simultaneous Interpreting: A Strategy or a By-product of Language Mediation? *Across Languages and Cultures* 7, 171–190. https://doi.org/10.1556/Acr.7.2006.2.2

Kajzer-Wietrzny, M., 2012. *Interpreting universals and interpreting style* (PhD Thesis). Adam Mickiewicz University, Poznań.

Magnifico, C., Defrancq, B., 2020. Norms and gender in simultaneous interpreting: A study of connective markers. *Translation & Interpreting* 12, 1–17. https://doi.org/10.12807/ti.112201.2020.a01

Van Besien, F., Meuleman, C., 2008. Style Differences among Simultaneous Interpreters: A Pilot Study. *The Translator* 14, 135–155. https://doi.org/10.1080/13556509.2008.10799252

Translating from Chinese into English: How can we learn more?

Yi Gu, Ana Frankenberg-Garcia
University of Surrey
y.gu@surrey.ac.uk, a.frankenberg-garcia@surrey.ac.uk

Until not very long ago there was no real translation industry in China (Translators Association of China, TAC). Practically only political essays, government documents and some literature aimed at portraying the People's Republic of China to the outside world were translated into foreign languages by official government bodies (Yang 1999). Recently, with China becoming an increasingly global player, the need for translating non-literary and non-political texts out of Chinese has increased dramatically, particularly into English (TAC). Because few people outside China have sufficient knowledge of Mandarin to be able to translate the language, most translation activity is carried out into the translator's second language (L2), also known as inverse translation. Yet little is actually known about texts translated from Chinese into English. Although there are several corpus studies examining Chinese translation (e.g., McEnery et al 1999, Xiao 2010, Xiao & Hu 2015), most such studies are based on unidirectional English into Chinese parallel corpora. In view of the dearth of corpus-based studies focusing on Chinese into English translation, this paper poses the following research questions:

i. What Chinese-English parallel corpora are available?
ii. What are the challenges of compiling a parallel corpus representative of current Chinese to English translation activity?
iii. How can we improve our understanding of Chinese into English translation?

To answer to the first of the above questions - what Chinese-English parallel corpora are available - we carried out a thorough review of existing parallel corpora catering for English and Chinese, discussing the text types they contain, their dates of publication, the translation language direction, the size of the corpora and how they can be accessed by researchers in general. We concluded that although there are reasonable amounts of parallel text data for Chinese and English, most of the corpora available are in the English into Chinese translation direction, or the exact translation direction is not specified or controlled for. The corpus resources available in the Chinese to English translation direction are few, scattered and limited.

The second research question - what are the challenges of compiling a parallel corpus representative of current Chinese to English translation activity - led us to conduct an extensive search for Chinese source texts translated into English from the 1990s onwards that are currently available online and that can be downloaded for the compilation of a parallel corpus. The first and most significant challenge we encountered was finding sources other than official government texts and public speeches, and finding English target texts that are full translations, rather than summary translations or heavily edited versions of texts aimed at foreign readerships. The second challenge was finding a parallel text aligner that worked well for Chinese and English. We tested nine aligners that support Chinese and English and compared them in terms of cost, access, performance, alignment output and alignment editing interface. The third challenge was how to address in a parallel corpus the notable text expansions and reductions that seem to be a common practice in Chinese to English translation, a matter which has non-trivial implications for parallel text alignment. Our discussion of these challenges can be useful for any researcher wishing to develop a parallel corpus of contemporary Chinese to English translation.

To answer the third question - how can we improve our understanding of Chinese into English translation - we have compiled ZHEN, a corpus of circa one million characters of contemporary simplified Chinese source texts (ZH) aligned with authentic translations into English (EN). Unlike existing parallel corpora in

this language direction, ZHEN is made up of a wide range of text types. Apart from the more readily available government reports, white papers, legal documents, public speeches and United Nations documents, the corpus also contains contemporary Chinese literature and film subtitles, bilingual texts from Chinese business, culture and university websites, and research abstracts which have been translated into English. At the moment of writing this abstract, ZHEN has been fully aligned and compiled, with 806,986 tokens in Chinese and 1,003,375 tokens in English. ZHEN has been compiled using Sketch Engine (Kilgarriff et al. 2014), a highly sophisticated yet user-friendly corpus platform that enables corpora to be shared online.

Unlike corpus-like parallel text browsers like Linguee and parallel corpora which do not differentiate between source texts and translations, or parallel corpora with English source texts and Chinese translations, ZHEN has been specifically compiled to enable one to understand directional shifts in Chinese to English translation. For instance, parallel concordances from Chinese to English can be used to analyse how Chinese culture-specific items have been translated into English, and countless other phenomena.

Additionally, the Chinese source texts of ZHEN can be contrasted with comparable corpora of (untranslated) Chinese, and the English translations of ZHEN can be contrasted with comparable corpora of (untranslated) English, in what Granger and Lefer (2020) refer to as the two-pronged approach to corpus-based crosslinguistic studies. A preliminary investigation contrasting adverb use in ZHEN and with adverbs in the English enTenTen13 and the Chinese zhTenTen17 corpora on Sketch Engine, for example, unveils significant differences in the use of epistemic stance adverbials in ZHEN. Our findings suggest that the language that is exported from China expresses a much higher degree of certainty than untranslated Chinese or English.

In conclusion, ZHEN can be useful not only to Translation Studies scholars, but also as training data in Machine Translation research, as a pedagogical instrument for translation tutors and students, and as a reference for professional translators.

References

Granger, Sylviane, and Lefer, Marie-Aude. 2020. “Introduction: A two-pronged approach to corpus-based crosslinguistic studies”. *Languages in Contrast* 20 (2): 167-183.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. “The Sketch Engine: Ten Years On.” *Lexicography* 1 (1): 7–36.

McEnery, Tony, and Richard Xiao. 1999. "Domains, text types, aspect marking and English-Chinese translation." *Languages in Contrast* 2 (2): 211-229.

Translators Association of China. [online] Available at: http://tac-online.org.cn/en/ [Accessed 20 Feb. 2020].

Xiao, Richard. 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.

Xiao, Richard, and Hu Xianyao. 2015. *Corpus-based studies of translational Chinese in English-Chinese translation*. Springer Berlin Heidelberg.

Yang, Zhengquan. 1999. “Xu [Introduction].” In *Zhongguo waiwenju wushinian dashiji [A chronology of China Foreign Languages Publishing and Distribution Administration (1949-1999)]*, edited by Yannian Dai and Rinong Chen. Beijing: New Star Press, I-VI.

Constructional transfer: compound constructions in second language acquisition

Isa Hendrikx[1], Kristel Van Goethem[2]
Université de Liège[1], F.R.S.-FNRS Université[2]
isa.hendrikx@uliege.be[1], kristel.vangoethem@uclouvain.be[2]

In this contribution we will present the objectives and methodology of a research project on the effects of cross-linguistic influence in Second Language Acquisition (SLA) and translation. Specifically, the project focuses on the use of Dutch compound constructions by French-speaking learners of Dutch and novice translators (L1 Dutch).

Languages are known to vary significantly with respect to their preferences for analytic or synthetic constructions (Rainer 2015). For instance, Germanic languages tend to use compounds more frequently than Romance languages (Van Goethem 2009; Schlücker 2019). This cross-linguistic variation has been described at length for Germanic (e.g. Hüning et al. 2006) and Romance languages (e.g. Lamiroy 2011). Nevertheless, little attention has been paid so far to the impact of these different cross-linguistic tendencies on the use of compounds in translation and in SLA, even though word-formation awareness has been proven crucial for learners' second language (L2) proficiency and creativity (Balteiro 2011), and constitutes an important factor in producing target-like translations (Lefer 2012).

Based on the state-of-the-art literature, we assume that the written output of both French-speaking L2 learners of Dutch and novice students translating French into Dutch will undergo transfer/priming effects from French, their L1/source language. In the domain of compounding, this would result in an underrepresentation of compounds, compensated by an overrepresentation of equivalent syntactic constructions in L2/translated Dutch, when compared to native/non-translated Dutch. However, the transfer effect in SLA is expected to be more significant than the priming effect in translation, since the L1 Dutch translators can already rely on a complete cognitive network of L1 compound constructions, while this is still under development in the case of L2 learners. In sum, the following cline with respect to the frequency of use of compound constructions can be hypothesized: L1/non-translated Dutch > translated Dutch > L2 Dutch > L1 French.

The project includes four corpus studies drawing on data from the *Multilingual Traditional Immersion and Native Corpus* (MulTINCo) (Meunier et al. 2020) and new corpus data.

The research project will extend the recent model of *Diasystematic Construction Grammar* (DCxG) to the study of SLA and translation. DCxG integrates constructionist insights into the study of multilingualism and conceptualizes the linguistic competence of multilingual speakers as an "interlingual network of constructions with different degrees of schematicity" (Höder 2012: 255). This so-called 'diasystem' contains "some constructions [that] are unspecified for language (such as abstract syntactic constructions and lexical concepts), while others (above all lexically and phonologically filled constructions) are language-specific" (Höder 2012: 247). Although Höder's framework has mainly been applied to issues of language contact, we believe it is highly relevant to SLA and translation studies: the understanding that in language production a bilingual's two languages are simultaneously activated and often processed in parallel engenders a "more organic view of bilingual cognition and the translation task" (Halverson 2014: 133).

In sum, embedding the results into the DCxG framework will enable us to compare the cognitive restructuring processes taking place in the minds of young multilingual learners and translators and to assess the role of constructional transfer and priming in foreign language learning and translation.

References

Balteiro, I. (2011). Awareness of L1 and L2 word-formation mechanisms for the development of a more autonomous L2 learner. *Porta Linguarum, 15.* 25-34.

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

Halverson S.L. (2014) Reorienting Translation Studies: Cognitive Approaches and the Centrality of the Translator. In: House J. (eds) Translation: A Multidisciplinary Approach. Palgrave Advances in Language and Linguistics. Palgrave Macmillan, London.

Höder, S. (2012). Multilingual constructions: a diasystematic approach to common structures. In K. Braunmüller & C. Gabriel (Eds.), *Multilingual individuals and multilingual societies.* Amsterdam/ Philadelphia: Benjamins. 241-257.

Hüning, M., Vogl, U., Van der Wouden, T., & Verhagen, A. (2006). *Nederlands tussen Duits en Engels. Handelingen van de workshop aan de Freie Universität Berlin*. Leiden: Stichting Neerlandistiek Leiden.

Lamiroy, B. (2011). Degrés de grammaticalisation à travers les langues de même famille. *Mémoires de la Société de linguistique de Paris, 19,* 167-192.

Lefer, M.-A. (2012). Word-formation in Translated Language: The impact of Language-pair Specific Features and Genre Variation. *Across Languages and Cultures, 13*(2). 145–172.

Meunier, F., Hendrikx, I., Bulon, A., Van Goethem, K. & Naets, H. (2020). MulTINCo: Multilingual Traditional Immersion and Native Corpus. Better-documented multi-literacy practices for more refined SLA studies. In L. Van Mensel & Ph. Hiligsmann (eds.) *Assessing CLIL: A multidisciplinary approach [special issue]. Journal of Bilingual Education and Bilingualism.* DOI: 10.1080/13670050.2020.1786494

Rainer, Franz. (2015). Intensification. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, & Franz Rainer (eds.), *Word-formation: An international handbook of the languages of Europe,* 1340–1351. Berlin/Boston: De Gruyter Mouton.

Schlücker, B. (Ed.). (2019). *Complex lexical units: compounds and multi-word expressions*. Berlin: De Gruyter.

Van Goethem, K. (2009). Choosing between A+N compounds and lexicalised A+N phrases: The position of French in comparison to Germanic languages. *Word Structure, 2,* 241-253.

Norms, constraints, risks: A usage-based perspective on sociocognitive constructs in corpus-based translation studies (and beyond)

Haidee Kotze[1], Sandra L. Halverson[2]
Utrecht University[1], University of Agder[2]
h.kotze@uu.nl, sandra.l.halverson@uia.no

Much corpus-based research on translation views the translated text as a product that carries linguistic or discursive traces of the sociocognitive processes "behind" it. Multiple sociocognitive processes are involved in the production of any given published translation (or any other text). It is not only the case that the text may be the result of more than one translator's work (or may have been shaped by the use of a translation memory, in itself an aggregate "repository" of human translational decision-making), but any published text also carries the traces of the work of other text producers, like revisers or copyeditors. All of this takes place within particular production settings, themselves embedded in broader cultural contexts. A translated text, therefore, is a palimpsestic record of the overlaid textual production processes of various people, shaped by and reflecting cognitive, interpersonal and contextual forces.

In attempting to capture this sociocognitive complexity of translation as a highly contextualised and embodied individual and intersubjective choice-making process, a range of explanatory constructs have been proposed. These constructs, which are widely invoked in corpus-based studies of translation, all acknowledge that there is an interplay between the individual and the collective, and the cognitive and the social, but nevertheless tend to place more emphasis on one dimension than the other. These constructs thus range from those with a more "social" emphasis to those with a more "cognitive" emphasis – expressed, in Malmkjær's (2005) terms, as an explanatory tension beteen "norms" and "nature" (see Kotze 2019 for further discussion). Towards the more "social" end of the continuum, particularly influential have been the "norm" construct (as developed for example in Toury 1995, 2012 and Chesterman 1993), and the notion of "risk avoidance" (as developed for example in Pym 2015). More explicitly crossing the boundary between cognitive and social accounts is the notion of "constraints" (as developed in the work of Kotze and Van Rooy; see Kruger & Van Rooy 2016; Kotze 2019, in press), based on earlier ideas by Baker (1993) and Lanstyák and Heltai (2012).

In this paper, we focus specifically on a critical evaluation of "norms" and related constructs as these are used in translation studies. They all, in one way or another, express the idea that the cognitive processes of translational decision-making are influenced by social experience with translation, and with language more generally, in real communicative contexts. However, the exact nature of the interaction between the cognitive and the social is not always clearly conceptualised – a point also raised by others. Meylaerts (2008), for example, attempts to resolve the tension by recourse to a sociological approach that focuses on how socialisation establishes norms, while Robinson (2020) proposes an alternative view of translational norms, based on 4EA cognition, that takes account of norm theory as developed in psychology and cognitive science.

The argument proceeds in three main steps. The first part of the paper focuses on the inherently sociocognitive nature of translational norms, and asks to what degree existing norm-related theoretical constructs in translation studies (and the way they have been interpreted by scholars) highlight and theorise the complex relationship between these two dimensions of norm formation. This part of the paper relies on a close critical reading of key texts in which these concepts are proposed and explained. In the second part of the paper we engage more closely with the arguments of Robinson (2020). In particular, we highlight what we consider the benefits and limitations of the 4EA paradigm as an explanatory framework. We follow this, in the third part of the argument, with a proposal for reconceptualising norms

as a sociocognitive construct that is founded on 4EA cognition, but that *also* takes due account of the linguistic nature of translation (see Halverson 2021). Specifically, we argue that any picture of translational decision-making is incomplete without considering linguistic accounts of normativity as both a bottom-up linguistic conventionalisation process (developing through experience), and a top-down linguistic legitimisation process (likewise developing through experience). Thus, we propose that normativity needs to be seen as both conventionalisation and legitimisation, and that these two forces are in constant interaction, through experience with language in everyday communicative settings (see also Kruger and Van Rooy 2017 for similar arguments in a different context).

To theorise this complexity, what is needed is a theory of language that explicitly considers how social experience with language leads to (cognitive) linguistic representations that carry within them the social embeddedness of language, including normative associations – which, in turn, prompt their use in particular contexts, either consciously or unconsciously, setting into motion a conventionalisation-legitimisation feedback loop. Usage-based linguistics provides such a framework, also amenable to the 4EA paradigm (see for example Bybee 2010 in general; Backus & Spotti 2012; Harder 2012 specifically in relation to a usage-based perspective on norms). We show how experience with language, for translators as much as for other language users, is the driving force behind the interplay of norms-as-convention and norms-as-legitimisation, and argue that a usage-based framework provides the theoretical apparatus to conceptualise the cognitive and social, and individual and collective, nature of translational norms and their development, within the broader frame of 4EA cognition. This leads to a destabilisation of the conceptual opposition between "norms" and "nature", cognition and society, and individual and collective. In conclusion, we reflect on the implications of this proposal for corpus-based studies of translation, across various areas of investigation.

References

Backus, A. & M. Spotti 2012. Normativity and change: Introduction to the special issue on *Agency and power in multilingual discourse*. *Sociolinguistic Studies* 6(2): 185-208.

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Chesterman, A. 1993. From "is" to "ought": Laws, norms and strategies in Translation Studies. *Target* 5(1): 1-20.

Halverson, S. 2021. Translation, linguistic commitment and cognition. In F. Alves & A.L. Jakobsen, eds. *The Routledge Handbook of Translation and Cognition*. London: Routledge. pp. 37-51.

Harder, P. 2012. Variation, structure and norms. *Review of Cognitive Linguistics,* 10(2): 294-314.

Kotze, H. 2019. Converging *what* and *how* to find out *why*: An outlook on empirical translation studies. In L. Vandevoorde, J. Daems & B. Defranq, eds. *New Empirical Perspectives on Translation and Interpreting* (Routledge Advances in Translation and Interpreting Studies). London: Routledge. pp. 333-371.

Kotze, H. In press. Translation as constrained communication: A multivarietal, multivariate approach. In S. Granger & M.-A. Lefer, eds. *Extending the Scope of Corpus-based Translation Studies*. London: Bloomsbury.

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

Kruger, H. & B. van Rooy. 2017. Editorial practice and the progressive in Black South African English. *World Englishes*, 36(1): 20-41.

Lanstyák, I. & P. Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1): 99-121.

Malmkjær, K. 2005. Norms and nature in Translation Studies. *SYNAPS* 16: 13-19.

Meylaerts, R. 2008. Translators and (their) norms: Towards a sociological construction of the individual. In M. Shlesinger, D. Simeoni & A. Pym, eds. *Beyond Description Translation Studies: Investigations in Homage to Gideon Toury*. Amsterdam: John Benjamins. pp. 91-102.

Pym, A. 2015. Translating as risk management. *Journal of Pragmatics*, 85: 67-80.

Robinson, D. 2020. Reframing translational norm theory through 4EA cognition. *Translation, Cognition & Behavior*, 3(1): 122-142.

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

Toury, G. 2012. *Descriptive Translation Studies – and Beyond* (revised edition). Amsterdam: John Benjamins.

Simplification and interference in English-Hungarian and Hungarian-English translated and interpreted texts in the EPTIC inter-modal sub-corpus

Timea Kovács
Karoli Gaspar University of the Reformed Church
kovacs.timea.phd@gmail.com

The significance of Corpus Linguistics in Translation Studies was first emphasised by Mona Baker in her two seminal papers in 1993 and 1995. A few years later, in 1998, Shlesinger highlighted the relevance of research based on Corpus Linguistics in Translation Studies. Since then, there have been varied attempts aimed at compiling and doing research in monolingual or multi-lingual, parallel and comparable corpuses. Nevertheless, relatively few inter-modal corpuses, incorporating written and spoken texts and their interpreted and translated counterparts, have been devised which can be used as mono- or multi-lingual parallel and comparable corpuses as well. The development of the EPTIC corpus was started by Bernardini et al. with a view to filling up this gap in the field of Corpus Linguistics. The aim of the EPTIC project is to examine and compare lexical simplification, introduced by Laviosa, in translated and interpreted texts in different language pairs and directions.

The aim of this paper is to introduce the EPTIC project and illustrate how lexical simplification functions in the process of English-Hungarian translation and interpreting through the micro-analysis of aligned English–Hungarian translated and interpreted texts taken from the above corpus.

Lexical simplification is also characteristic of English–Hungarian language mediation. Heltai (2002) claims that the vocabulary of translated texts is more limited and the average sentence length is less than that of untranslated texts. According to Chesterman's theory (2004), the source language always leaves a mark on the target language, along with the target language universal that translated texts are always simpler than untranslated texts. Nevertheless, in line with Toury's theory of interference (1995), interference in the process of language transmission can result in grammatical and lexical explications, i.e. grammatical and lexical level explanations, extensions, and more complex structures. This seems to contradict the principle of simplification as a universal translation.

As the creation and research of Hungarian language interpreter corpus was somewhat relegated to the background compared to translation, I have relied on the methods and relevant data of the EPTIC inter-modal corpus. During the research in the English–Hungarian language EPTIC sub-intermodal corpus, I am looking for the answers to the following questions:

(1) Is the text translated from English into Hungarian lexically simpler than the translated one?
(2) Is interference more noticeable during translation or interpretation?
(3) Does interpretation lead to the simplification of content?

On the basis of the results of the research conducted (so far), it can be stated that the sentences in the text interpreted into Hungarian are much shorter than in the translated text. As sentence length is a feature of lexical simplification, it can be concluded that the Hungarian-English interpreted text is lexically simpler than its translated counterpart.

As for interference, the negative effect (interference) of the analytical source (English) language on the synthetic target (Hungarian) language has been examined on the basis of the frequency of the functional verb structures. The examination is based on the assumption that one of the measurable results of source-language English interference is the use of analytical (functional verb) structures versus the use of verbs typical of Hungarian (synthetic) morphosyntactics. As a result of the analysis, it can be observed that

functional verb structures occur in a much higher proportion in the translated text, probably as a result of interference, than in the untranslated Hungarian texts. Nevertheless, the proportion of function verb structures in the interpreted text is much lower. This can be considered as an example of grammatical implication, which may be the result of the fact that Toury's rule of source language interference (the effect of analytical structures) is less prevalent in the oral (interpreted) text, which is more characterised by the use of verbs characteristic of Hungarian (synthetic) spoken language morphosyntactics. It is a question of whether the different incidence rates of functional verb structures are indeed the result of grammatical implication (simplification), which is more characteristic of oral mediation, or of the weaker effect of source language interference.

With regard to the relation between interpretation and simplification, the content elements in the mediated (interpreted into Hungarian) and the untranslated (English speeches) were compared. Based on the results of the comparison, significant differences can be seen in the prevalence of content elements. These can be partly explained by the different morphosyntactic and syntactic structures of the two languages. In some segments, a given content element is inserted, which may function as a cohesion element. However, it is more common to omit a content element from the interpreted text, which can even change the exact message of that segment. Therefore, the omission and random insertion of content elements of the source language text in the interpreted target language text may result in content simplification (loss and modification).

On the basis of the above results, questions worthy of further research are formulated: does the interpreted text converge better with the spontaneously spoken linguistic Hungarian, or is the source language (English) interference stronger in the case of the translated text? Do other linguistic strategies adopted in simultaneous interpretation result in the simplification of content? Confirming the results of the research so far and answering the above questions require the continuation of the research and its extension to further language directions (Hungarian-English).

References:

BAKER, Mona: Corpora in Translation Studies. An Overview and Suggestions for Future Research, *Target*, 7, 2, 1995, 223–245.

BAKER, Mona: Corpus Linguistics and Translation Studies. Implications and Applications, in Mona Baker – Gill Francis – Elena Tognini-Bonelli (eds.): *Text and Technology: In Honour of John Sinclair*, Amsterdam, Benjamins, 1993, 233–250.

BERNARDINI, Silvia – FERRARESI, Adriano – MILICEVIC, Maja: From EPIC to EPTIC — Exploring Simplification in Interpreting and Translation from an Intermodal Perspective, *Target*, 28, 1, 2016, 61–86.

CHESTERMAN, Andrew: Hypotheses about Translation Universals, in Gyde Hansen – Kirsten Malmkjær – Daniel Gile (eds.): *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*, Benjamins Translation Library, 2004, 1–13.

HELTAI, Pál: "Claims, Changes and Challenges". Az EST III. Nemzetközi Kongresszusa. Koppenhága, 2001. augusztus 30 – szeptember 1., *Fordítástudomány*, 4, 1, 2002, 123–133.

KAJZER-WIETRZNY, Marta: *Interpreting Universals and Interpreting Style*, Poznan, Adam Mickiewicz University, 2012, [Doctoral dissertation].

LAVIOSA-BRAITHWAITE, Sara: *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*, Manchester, UMIST, 1996, [PhD Thesis].

LAVIOSA, Sara (ed.): Special Issue – The Corpus-based Approach: A New Paradigm in Translation Studies, *Meta*, 43, 4, 1998a.

LAVIOSA, Sara: Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose, *Meta*, 43, 4, 1998b, 557–570.

RUSSO, Mariachiara – BENDAZZOLI, Claudio – SANDRELLI, Annalisa: Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended Analysis of EPIC, *Forum*, 4, 1, 2006, 221–254.

SHLESINGER, Miriam – ORDAN, Noam: More Spoken or More Translated? Exploring a Known Unkown of Simultaneous Interpreting, *Target*, 24, 1, 2012, 43–60.

SHLESINGER, Miriam: Towards a Definition of Interpretese: An Intermodal, Corpus-based Study, in Gyde Hansen – Andrew Chesterman – Heidrun Gerzymisch-Arbogast (eds.): *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile*, Amsterdam, John Benjamins, 2009, 237–253.

SHLESINGER, Miriam: Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies, *Meta*, 43, 4, 1998, 1–8.
TOURY, Gideon: *Descriptive Translation Studies and Beyond*, Amsterdam and Philadelphia, John Benjamins, 1995.

Equivalent or Effective? – Correlating Speech Register Variation with Audience Perception of the Conference Interpreting Product

Nannan Liu
The University of Hong Kong
nnl93@connect.hku.hk

This research examines whether the pragmatic effect of source and interpreted speeches in political conference interpreting is equivalent to each other. Pragmatic effect refers to audience members' attitudinal and interpersonal response (Pöchhacker 2016: 145–148). Whereas such a type of effect has been addressed extensively via register shift research in court interpreting, conference interpreting scholars have restricted their analyses to audience perception of interpreter intonation and accent (Collados Ais 2002; Cheung 2013). In the present research, I adopt the construct "register" (termed "speech type" for laypersons in the perception study) as the operationalisation of such a type of effect, with register defined as a variety of spoken language "associated with a particular situation of use" to serve particular communication purposes (Biber & Conrad 2019: 6).

This study relies on a self-created speech corpus of 358 consecutive interpreting (CI) sequences and a speech type survey instrument. The speakers are two Chinese premiers and interpreters are six government staff working into their B language. It examines the speech register of source and target utterances (SUs and TUs) with a cross-linguistic multidimensional (MD) approach (Liu 2021) and an audio classification task. The survey instrument comprises 26 items gathered based on a review of perceptual descriptors of the interpreted speech (e.g., Collados Ais 2002; Baigorri-Jalon 2014). It was administered to 1892 US residents, with eighty per cent of interpreting audio segments rated three to six times. Linear regressions with predictor selection were conducted between registerial, prosodic, and topical predictors and the response variable pragmatic reaction.

Register shifts between SUs and TUs were directly quantified with hierarchical agglomerative clustering (Liu 2021). Linguistically, the interpreters concerned are demonstrated to have shifted literate source to a more oral, attitudinal, and less formal register and oral source to a less oral, more attitudinal, and formal register. Regardless of source types, the interpreting products tend to be more similar to each other than SU registers. Prosodically, the *sui generis* "interpreter's intonation" attested in simultaneous interpreting research (Lenglet & Michaux 2020) was found to exist in CI as well. The most salient discriminators of SUs versus TUs were stresses per second, intonation variability, and pauses. The interpreters concerned tended to stress every word in a flat tone with few pauses. In contrast, the source speech is marked by prosodic plurality: literate SUs exhibited patterns emblematic of informal speech whereas oral SUs were delivered in a fast pace and monotone. Such multiplicity challenges the assumption that the "formal settings" of conference interpreting necessarily involve formal source speech (AIIC 2020).

Reliability tests of survey items and audio segments were conducted, with the results indicating 24 items (such as "detached–interactive", "monotonous–varied", and "unclear–clear") and 302 segments to elicit consistent reactions from survey respondents. The respondents were balanced in gender and most had a bachelor's degree, with a mean age of 37.47 years. ANOVA tests demonstrated age and gender to have significant, albeit limited, effects on the pragmatic perception. The majority of respondents uniformly perceived TU audio segments as professional, strong, unemotional, monotonous, detached, and unexciting. My investigation into the polar ratings revealed topics of the segments to be overarching factors conditioning the pragmatic effect.

Three pragmatic dimensions were gleaned from survey responses. TU segments that were rated and representative entered correlation analyses. The results indicated interpreters' performance on a dimension of register variation, i.e., "informational elaboration under real-time production constraints" (Biber 1988: 154), to correlate significantly with the perception of the lack of engagement with the audience. Such a dimension and the paucity of loudness variability in TUs are also correlated with the negative perception of interpreters' delivery.

In sum, this research reports that by any measure, be it linguistic, paralinguistic, or perceptual, the pragmatic impact of interpreting is not equivalent to that of the original speech. The conference setting was found to host a wide range of source speech types, prosodic strategies, and audience expectations, but interpreters tended to ignore such multiplicity and convey an "equalised" (Shlesinger 1989) range of pragmatic effect. The extent to which such results were constrained by language specificity, interpreting-inherent challenges, target register conventions, and meta- and institutional discourse was discussed. In alliance with previous studies (e.g., Diriker 2004), this research questions an unexamined belief in "transcendental signified" and "equivalence of effect".

References

AIIC. (2020). Conference Interpreting [Professional organisation]. Retrieved August 6, 2020, from aiic.net website: https://aiic.org/site/world/conference?

Baigorri-Jalon, J. (2014). *From Paris to Nuremberg: The Birth of Conference Interpreting* (H. Mikkelson & B. S. Olsen, Trans.). Amsterdam: John Benjamins.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (Second edition). Cambridge: Cambridge University Press.

Cheung, A. K. F. (2013). Non-native accents and simultaneous interpreting quality perceptions. *Interpreting*, *15*(1), 25–47. https://doi.org/10.1075/intp.15.1.02che

Collados Ais, A. (2002). Quality Assessment in Simultaneous Interpreting: The Importance of Nonverbal Communication. In F. Pöchhacker & M. Shlesinger (Eds.), *The Interpreting Studies Reader* (pp. 326--336). London; New York: Routledge.

Diriker, E. (2004). *De-/Re-Contextualizing Conference Interpreting*. Amsterdam: John Benjamins.

Lenglet, C., & Michaux, C. (2020). The impact of simultaneous-interpreting prosody on comprehension: An experiment. *Interpreting*, *22*(1), 1–34.

Liu, N. (2021). Register shifts in political conference interpreting: A multidimensional analysis. In J. Pan, S. L. Halverson, & J. Munday (Eds.), *Translating and Interpreting Political Discourse: New Trends and Perspectives*. Leiden: Brill.

Pöchhacker, F. (2016). *Introducing Interpreting Studies* (2nd ed.). London and New York: Routledge.

Shlesinger, M. (1989). *Simultaneous interpretation as a factor in effecting shifts in the position of texts in the oral-literate continuum* (MA Thesis). Tel Aviv University, Tel Aviv.

The Use of Italian and French Oppositional Connectors and their Translation to Lithuanian in the Opinions of the Advocate General in EUR-Lex

Jurgita Macijauskaitė-Bonda, Aurelija Leonavičienė
Vytautas Magnus University
jurgita.macijauskaite-bonda@vdu.lt, aurelija.leonaviciene@vdu.lt

Since 1990, the year of the re-establishment of the independence of the Republic of Lithuania, and especially after 2004 when the country entered the European Union, a significant development of administrative Lithuanian began. It was fostered by translators of administrative texts and, in particular, Lithuanian Translation Units of The Directorate-General for Translation of the European Commission and the European Parliament as well as other translation units of the EU.

The present paper analyses a specially compiled corpus of the Opinions of the Advocate General of the Court of Justice of the EU written in Italian and French and translated into Lithuanian in the years 2018-2020 available in the EUR-Lex database. The comparative corpus covers more than 300 000 words of textual material of legal discourse. The object of the research is to analyse oppositional connectors common to French and Italian legal texts extracted from the comparative corpus. Oppositional connectors help to impart oppositional relations between different text components and to create a coherent and logical text.

In order to conduct a comparative analysis on the use of oppositional connectors in French (mais, pourtant, néanmoins, etc.) and Italian (ma, però, contrariamente a, etc.) and their translation to Lithuanian, the data collected from the EUR-Lex database was converted into the textual format (txt.) and further processed by text analysis software AntConc 3.5.7 (Anthony 2018) which helped to extract word frequency lists and concordances, providing useful information on the contextual use of oppositional connectors. The total number of oppositional connectors that were found in the analysed administrative texts in Italian and French is around 350 for each language. Research also aims to holistically evaluate the significance of the use and functions of connectors, determine the quantitative distribution of translation techniques as well as reveal general tendencies in the translation of oppositional connectors from the analytic languages, namely, Italian and French to Lithuanian – a synthetic language.

The analysis of the distribution and contextual use of oppositional connectors is based on the scholarly works of Italian and French linguists (Visconti 2011; Dardano, Trifone 2009; Adam 2005; Roulet 1999, etc.), whereas research into the translation of oppositional connectors from Italian and French to Lithuanian is based on the theoretical and methodological insights of corpus linguists (Frœliger, 2013; Marchand, 1998; Habert, Nazarenko, Salem 1997, etc.). For the research into the distribution and translation of oppositional connectors, the methods of corpus linguistics, statistical and comparative analysis as well as descriptive-analytical approach are applied.

The analysis of word frequency and concordance lines shows that French oppositional connectors comprise around 36,4 % of all connectors that occurred in the data and around 35,2 % in the Italian part of the corpus. The statistical distribution of oppositional connectors clearly indicates the standardised use of connectors, their frequent repetition and little synonymy in conveying opposition between parts of the text. The analysis of the paralleled French-Italian-Lithuanian data (AntPConc, Anthony 2018 a) revealed prevalent translation techniques used to translate Italian and French oppositional connectors to Lithuanian, namely, 1) direct translation, 2) selecting contextual synonyms, and 3) omission.

References

Adam, J. M. (2005). *Linguistique textuelle. Des genres de discours aux textes*. Une introduction méthodique à l'analyse textuelle des discours. Paris: Nathan.

Anthony, L. (2018). AntConc (3.5.17) [Windows]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Anthony, L. (2018 a). AntConc (1.2.1) [Windows]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Bertels, A., De Hertog, D., Haylen, K. Étude sémantique des mots-clés et des marqueurs lexicaux stables dans un corpus technique. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2: TALN, pages 239–252, Grenoble, 4 au 8 juin 2012. https://www.researchgate.net/publication/263806866_Etude_semantique_des_mots-cles_et_des_marqueurs_lexicaux_stables_dans_un_corpus_technique

Bolzoni, L. (2008). Elementi pragmatici nel testo scientifico: un'analisi contrastiva fra connettivi italiani e francesi. In: Ursula Reutner, Schwarze Sabine (ed.). *Le style, c'est l'homme: Unité et pluralité du discours scientifique dans les langues romanes*, p. 227–248. Frankfurt am Main: Peter Lang.

Dardano, M, Trifone, P. (2009). *Grammatica italiana. Con nozioni di linguistica*. Bologna: Zanichelli. Terza edizione.

Frœliger, N. (2013). *Les noces de l'analogique et du numérique. De la traduction pragmatique*. Paris: Les Belles Lettres.

Habert, B, Nazarenko, A, Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin/Masson.

Author 1, Rečiūnaitė, J. (2013). Prancūzų administracinio ir publicistinio stiliaus konektorių vartojimas ir vertimas į lietuvių kalbą. *Kalbų studijos / Studies about languages*, 22, p. 48–54.

Marchand, P. (1998). *L'analyse du discours assistée par ordinateur*. Paris: S.E.S.J.M./ Armand Colin.

Reboul, A., Moeschler, J. (1998). *Pragmatique du discours. De l'interprétation de l'énoncé à l'interprétation du discours*. Paris: Armand Colin.

Roulet, E. (1999). *La description de l'organisation du discours. Du dialogue au texte*. Paris: Didier.

Visconti, J. (2011).Tradurre i connettivi. Trattati e testi normativi. *CDCT Working Paper* n.2 / 2011 - 15 Novembre Cc, p. 2-12. http://www.cdct.it/wp-content/uploads/2014/06/116_1.pdf

Visconti, J. (2017). Riflessioni linguistiche sulla traduzione: il connettivo 'o' nelle sentenze della Corte di Giustizia dell'Unione Europea, *CERTEM*, Publifarum, n. 27. http://publifarum.farum.it/ezine_pdf.php?id=389

Visconti, J. (2000). *I connettivi condizionai complessi in italiano e in inglese. Uno studio contrastivo*. Torino: Edizioni dell'Orso.

Eur-lex material used in the work is © European Union, available at http://eur-lex.europa.eu/, 1998–2019.

Preparing parallel corpora for intralingual machine translation

Jean Nitzke[1], Silvia Hansen-Schirra[2], Silke Gutermuth[2]
University of Agder[1], University of Mainz[2],
jeann@uia.no, hansenss@uni-mainz.de, gutermsi@uni-mainz.de

Using corpora for machine translation (MT) has been inevitable since the data driven approaches to MT were introduced in the late 1980s. While the approaches to MT have changed and developed, most of them still draw on parallel corpus data. Hence, the requirements for the corpus data and how to prepare and align the data have been very similar ever since. However, when training MT for new translation tasks, e.g. for intralingual translation, the requirements for preparing the training corpora might change. The demand for translations into German Easy Language has been rising in the last couple of years, because society's awareness and the legal situation concerning accessible communication have changed in Germany. Easy Language is a variety of Standard German that adheres to specific rules, which simplify textual complexity and thus enhance comprehensibility. Usually, those texts are intralingual translations of standard texts. The main target group are people with cognitive disabilities, however other groups might benefit from Easy Language as well, like non-native German speakers (Bredel and Maaß 2016). In order to cope with the growing demands for Easy Language translation, we suggest that intralingual Neural Machine Translation (NMT) might be an option and want to discuss how to prepare parallel corpora as training data.

The combination of pre-translating a text by an MT system and post-editing (PE) the raw MT output by a professional translator or post-editor has been established as a workflow option for professional translations in the last ten to fifteen years (e.g. Porsil 2017 or Gaspari et. al 2015). Traditionally, MT and PE were thought to be most helpful for domain-specific texts with standardised source text language. In recent years, however, new text types have become of interest in MT and PE research like the translation of literature (e.g. Toral et. al 2018). Further, the quality of the NMT output is similar for text written in standard and controlled language (Marzouk and Hansen-Schirra 2019). Hence, training NMT systems for intralingual translation purposes seems interesting, as well, to provide an additional aid for professionals as the demands for those translations will be further rising in the coming years.

In contrast to interlingual translation, the translation from Standard German (SG) into Easy German (EG) requires varying translation strategies. Although strategies like additions and omissions are well known both in interlingual and intralingual translation, the amount in which they occur differs a lot for SG and EG. Another strategy is reduction of information which is a rather uncommon strategy in interlingual translation, but a common strategy for Easy Language to meet the demands of the main target group. Further, information are restructured and presented in a very different way. Hence, intralingual translation requires more than simple complexity reduction methods as known in controlled languages or in complexity reducing algorithms (as e.g. in Aluísio et. al. 2008 or Cheng et. al. 2016). As the quantity and quality of information delivered in the source and target text differ, the alignments are only very rarely 1:1 alignments, even on the sentence level. In addition to n:m alignments (1), we also encounter empty link alignments (2). Often, it is not obvious to which segments the information belong (3):

(1) SG: Bildung und Beschäftigung sind Schlüsselkomponenten für die Teilhabe an allen Bereichen unserer Gesellschaft. [Education and work are main components for participation in all parts of our society.]
EG: Alle Menschen sollen überall in unserer Gesellschaft dabei sein und mitmachen können. [All people should always be part of and interact with our society.]
Dafür müssen sie eine gute Schul-Bildung haben. [Therefore, they need a good edu-cation.]

Und sie müssen eine gute Arbeit haben. [And they need good jobs.] *(ST Arbeit)*

(2) SG: Berufliche Förderung [Job Support]
EG: Berufliche Förderung [Job Support] {b>Berufliche Förderung<b} bedeutet: [{b>Job Support<b} means:] Menschen mit Behinderung finden schwerer einen guten Arbeits-Platz. [People with disabilities have more difficulties to find a good job.]
Aber ein guter Arbeits-Platz ist wichtig für die Menschen. [But a good job is important for people.] (ST Ausbildung)

(3) SG: Die offizielle deutsche Übersetzung der UN-Behindertenrechtskonvention spricht von einem „integrativen Bildungssystem". [The official German translation of the UN Convention of the Rights of Persons with Disabilities talks about an "integrative education system".] Die englische Fassung, die zu den sogenannten authentischen Sprachfassungen gehört, dagegen von „inclusive education". [The English version, which belongs to the so-called authentic language versions, however, talks about "inclusive education".] EG: Bisher gab es in Deutschland an einigen Schulen schon Integration. [So far, there has already been integration at some schools in Germany.]
Einzelne Kinder und Jugendliche mit Behinderung lernten mit Kindern ohne Behinderung zusammen. [Some children and teenagers with disabilities learnt together with children without disabilities.]
Inklusion bedeutet aber: [Inclusion means, however:]
Die Schulen müssen für alle Kinder mit und ohne Behinderung gut sein. [The schools have to be good for all children with and without disabilities.]
(ST Bildung)

In this presentation, we want to talk about preparing corpus data for intralingual translation, (SG to EG). First, we will talk about the problems that we encounter when collecting corpus data. Within this context, the following questions will be discussed: Which text types are suitable? Are the texts biased by different rule sets? Do all translators use similar strategies for intralingual translation? Further, we will explore different examples of problematic alignments and different strategies that can be applied keeping in mind the final goal of training an NMT system. Finally, we address the questions whether certain corpus data need to be excluded for the training corpus of the NMT system and how translation problems that cannot be tackled by the NMT system have to be solved by professional post-editors.

References

Bredel, Ursula, and Christiane Maaß. (2016). Leichte Sprache: Theoretische Grundlagen? Orientierung Für Die Praxis. Bibliographisches Institut GmbH.

Gaspari, Federico, Hala Almaghout, and Stephen Doherty. (2015). 'A Survey of Machine Translation Competences: Insights for Translation Technology Educators and Practitioners'. *Perspectives* 23 (3): 333–358.

Marzouk, Shaimaa, and Silvia Hansen-Schirra. (2019) "Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures." *Machine Translation* 33.1-2: 179-203.

Porsiel, Jörg. (2017). Machine Translation - What Language Professionals Need to Know. Fachkommunikation. Berlin: BDÜ Fachverlag.

Toral, Antonio, Martijn Wieling, and Andy Way. (2018). 'Post-Editing Effort of a Novel with Statistical and Neural Machine Translation'. *Frontiers in Digital Humanities* 5: 9.

How do African Americans Speak in Finnish? The Translation of African American English into Finnish in Translated Finnish Fiction

Tomi Paakkinen
University of Turku, Finland
tomipaakkinen.tp@gmail.com

My PhD study is about what type of linguistic features Finnish translators have used to represent African American English. Under investigation are the lexical, morphological, phonological and syntactic features of colloquial Finnish and the statistical frequencies of occurrence of these linguistic features in six translations of novels written by six different African American authors. The purpose is to determine whether the translation of African American dialogue differs from the translation of spoken language in general in terms of the type and number of linguistic features used by the translators.

My presentation will first introduce the topic and then describe the research method. Then the presentation will move onto the first results obtained from the study, and how these results differ from previous studies.

Sampo Nevalainen (2004) observes, based on data extracted from the Corpus of Translated Finnish, a comparable corpus consisting of Finnish translations of different texts on the one hand, and of texts originally written in Finnish on the other hand, that Finnish translators prefer lexical features (colloquial words) to represent spoken language, whereas original Finnish authors prefer phonological features. Yet, the preliminary results of my PhD study may suggest that this is not always the case: the translators under study seem to prefer phonological features of speech rather than lexical, much like original Finnish authors. So far, the analysis of the translations of the novels *A Day Late and a Dollar Short* by Terry McMillan and *A Red Death* by Walter Mosley show the translators' clear preference for phonological features (in the former translation, 73% of all colloquialisms are phonological; in the latter, 63% are). This result corresponds to that of my previous study (Paakkinen 2013).

The research method has involved locating and quantifying features of colloquial Finnish using the concordance software AntConc. Although certain phonological and morphological features can be located using certain search parameters, locating and defining lexical features is more problematic, requiring some manual analysis. For example, finding instances of colloquial monophtongisation of diphthongs is a relatively straightforward procedure involving a wild card search (for example *ee*), but finding lexical features, such as slang words, is more problematic because it is difficult to predict the incidence of such features beforehand. Another problem is the arbitrariness of classifying a word as colloquial. In this study, words are defined as colloquial based on the descriptions given in the Dictionary of the Institute for the Languages of Finland (*Kielitoimiston sanakirja* 2017).

My data consist of the translations of six novels: *A Day Late and a Dollar Short* (2001) by Terry McMillan; *A Red Death* (1991) by Walter Mosley; *The Women of Brewster Place* (1982) by Gloria Naylor; *Fast Sam, Cool Clyde, and Stuff* (1975) by Walter Dean Myers; *If Beale Street Could Talk* (1974) by James Baldwin; and *The Spook who sat by the Door* (1969) by Sam Greenlee.

References

Kielitoimiston sanakirja. 2017. Institute for the languages of Finland. Available at: https://www.kielitoimistonsanakirja.fi/
Nevalainen, Sampo. 2004. Colloquialisms in translated text. Double illusion? *Across Languages and Cultures* 5, 67–88.
Paakkinen, T. 2013. A Study of African American Vernacular English in Three Novels and Colloquial Finnish in their Translations: *The Dark Tower II: The Drawing of the Three*, *A Time to Kill* and *Push*. Master's thesis, the University of Turku, English Translation and Interpreting.

Pragmatic strategies employed in the translation and interpreting of contrastive relations in political motion speeches: A corpus-based study

Jun Pan
Hong Kong Baptist University
janicepan@hkbu.edu.hk

Motions, referred as "any proposition brought before an assembly for its consideration" (Fielde, 1899, p. 13) constitute crucial components of any legislative and parliamentary procedures. As an official proposition, political motion speeches are usually delivered based on a written motion paper, followed by a debate or a question and answer session. In places where there is more than one official language, such papers are translated (and sometimes published) before the official assembly and delivered with simultaneous interpreting with text on site. Though sometimes symbolic, political motion speeches should be delivered and interpreted with high level of pragmatic competence so that they could be passed successfully and achieve targeted goals. In this aspect, contrastive relations consist of important rhetoric devices that should be dealt with care, especially in persuasive speeches delivered in political settings.

Contrastive relations are often textualized by the linguistic form of contrastive markers (CMs), which, often used as adversative conjunctions, can be illustrated by the use of expressions including *however*, *but*, *on the contrary*, etc. The use of CMs usually signals that "the utterance following is either a denial or a contrast of some proposition associated with the preceding discourse" (Fraser, 1996, p. 187), or with a subsequent utterance that is "contrary to expectation" (Halliday & Hasan, 1976, p. 250). They can even indicate "the unexpected, surprising natural of what is being said in view of what was said before" (Biber et al., 2002, p. 878).

Previous research has indicated that contrastive relations play important roles in both written and spoken texts (e.g. Malá, 2006; Taboada & Gómez-González, 2012). Pan and Wong (2018), in particular, identified a tendency to mitigate the contrastive relationship in Cantonese-English political interpreting. However, there is a lack of systematic research on the pragmatic strategies involved in the rendition of contrastive relationship in political speeches, and thus the development of relevant pedagogical measures lag far behind.

This paper, employing corpus-based approaches, aims to look into the pragmatic strategies employed in the translation and interpreting of contrastive relationship in political motion speeches. The language pair under investigation is Chinese (Cantonese) - English because of the remarkably pragmatic differences between these two languages (Pan and Wong, 2018).

The study tapped into data taken from two corpora, i.e., the Chinese/English Political Interpreting Corpus (CEPIC, Pan, 2019) and the Chinese/English Translation & Interpreting Learner Corpus (CETILC, under development): the former is an open-access 6.5 million wordtoken corpus on political interpreting, and the latter a learner corpus covering various topics and learner outputs at different learning stages. The Hong Kong Policy Addresses subset was taken from the CEPIC (1997-2017), which consists of Policy Addresses delivered by Chief Executives as motions at the Legislative Council and their translations and transcribed interpreted texts. The learner corpus subset used for comparison consists of translations and interpretations by students at advanced level of study. The source texts are motion papers of Children's Council in Hong Kong (which has been held annually since 2002). Both sub-corpora included a translation and interpreting subset: the latter transcribed with paralinguistic features including fillers, pauses, false starts, etc. Both sub-corpora were CM annotated and aligned at paragraph level. Since two CMs were found to be the most frequently used in both sub-corpora, i.e., "*bat gwo*" and "*daan (hai)*"

(corresponding to *however* and *but* in English), the analysis of translational strategies was focused on them.

The study further annotated and investigated the different strategies applied in translating and interpreting these two CMs, categorizing them as intensification, total equivalence, mitigation, and omission (cf. Kade, 1968; Martinovski, 2010). Contextual factors, such as semantic topic and disfluency indicators (e.g., pauses and fillers) were also explored.

Findings of the study show that both "*bat gwo*" and "*daan (hai)*" were translated into a variety of different contrastive markers in English. The study also indicates how contextual factors such as topics may influence the pragmatic strategies. For instance, "*bat gwo*" was mostly frequently rendered by professional translators and interpreters using the strategy of intensification when education served as the main topic, and the rest of the other three strategies (total equivalence, mitigation and omission) when economy was discussed. "*Daan (hai)*", however, where rendered most frequently by total equivalence, followed by mitigation and omission, and the topics usually concern economy and development. The comparison with learners suggests a somehow different pattern that can shed light on specific measures for pedagogic enhancement.

Acknowledgements
This study is supported by the Research and Development Projects (Standing Committee on Language Education and Research [EDB(LE)/P&R/EL/175/3]) and General Research Fund (Research Grants Council [12611717]).

References

Biber, D., Conrad, S., Reppen, R., Byrd, P., and Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36(1), 9–48.

Fielde, A. M. (1899). *Parliamentary Procedure: A Compendium of Its Rules Compiled from the Latest and Highest Authorities, for the Use of Students and for the Guidance of Officers and Members of Clubs, Societies, Boards, Committees, and All Deliberative Bodies*. Seattle/Washington: Helen N. Stevens.

Fraser, B. (1996). Pragmatic markers. *Pragmatics, 6*(2), 167–190.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Kade, Otto. (1968). *Zufall und Gesetzmässigkeit in der Übersetzung*. Leipzig: Verl. Enzyklopädie.

Malá, M. (2006). Contrastive markers and dialogicality. *Discourse and interaction*, *2*, 97-107.

Martinovski, Bilyana. (2010). Mitigation. In Louise Cummings (Ed.), *The Routledge Pragmatics Encyclopedia* (1st ed.), pp. 273-274. London & New York: Routledge.

Pan, J. (2019). The Chinese/English Political Interpreting Corpus (CEPIC). Hong Kong Baptist University Library, Accessed from https://digital.lib.hkbu.edu.hk/cepic/

Pan, J., & Wong, B. T. M. (2018). A corpus-driven study of contrastive markers in Cantonese– English political interpreting. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *9*(2), 168-176.

Taboada, M., & Gómez-González, M. (2012). Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences*, *6*(1-3), 17-41.

A comparative study of register feature changes in both translated and original Chinese texts based on Chinese Diachronic Composite Corpora

Shuangzi Pang
Shanghai Jiao Tong university
melody2459@hotmail.com

Previous research about language contact through translation (LCTT) has proved that translation, as one locus of language contact, plays a significant role in the history of language as it does in those of ideas and cultures. However, to date, the study about the role of translation on language change has been limited in three aspects. First, the corpus-based work in this domain has been inhabited by the availability of a large composite corpus covering different registers in genetically distinct language pairs. Second, previous research in this area has been confined into some individual features, and very little research considers the relationship between translated registers and non-translated registers in the same language. Third, the measurement of the impact of translated registers on the change of non-translated registers in Chinese has been constrained by the under-developed statistical approaches.

In order to address some of these gaps, we take the application of Multivariate Dimension (MD) method further in this article by comparing patterns of register variation in both translated Chinese and original Chinese across three sampling periods. The present study, from the perspective of English and Chinese, examines historical change of literary register features in both translated and original Chinese, and the interactive relationship between them based on the Chinese Diachronic Composite Corpus (CDCC). The aim of the investigation is to explore if translated registers in Chinese literary texts have an effect on those in original Chinese literary texts. This study thus attempts to answer the following questions:

(1) Do the register features in translated Chinese and original Chinese show significant change respectively across the three sampling periods? Are the register features caused by source language shining-through effect?
(2) Do the register features in translated Chinese show the same trend with those in original Chinese across the three periods? Are the register features in translated Chinese correlated with those in original Chinese over the three points?
(3) Do the individual features in each dimension concluded from MD analysis in translated Chinese impact on the equivalent dimension in original Chinese?

These questions will be answered with a newly established corpus called CDCC, which incorporates a diachronic English-Chinese parallel corpus, a comparable diachronic Chinese corpus and a “pure” Chinese reference corpus. The translated Chinese texts and original Chinese texts have been divided into three corpora, containing the texts from 1930s,1960s and 1990s.Based on the CDCC, this research reports on a multi-dimensional (MD) analysis of register variation in translated and original Chinese diachronically. The six dimensions identified by Biber (1988) and replicated in the analysis in the current study are:(1) involved versus informational production; (2) narrative versus non-narrative concerns; (3) explicit versus situation-dependent reference; (4) overt expression of persuasion; (5) abstract versus nonabstract information; and (6) online information elaboration.

A quantitative analysis at both the macro-level of factor (or dimension) scores and the micro-level of individual features provides substantive evidence for the effects of both processes of language change. In statistical analysis, ANOVA test is applied to test if each dimension in translated Chinese change significantly across three periods. Random forest is used respectively to measure which of the six dimensions contribute most substantially to a model predicting the period in which translated Chinese texts and original Chinese texts are produced.

Current findings demonstrate: (1) Translated Chinese texts demonstrate significant effects diachronically in Dimension 1 and Dimension 4. In contrast, original Chinese texts show significant change in each dimension across the three periods. Consequently, compared with original Chinese, translated Chinese keeps relatively stable across time. Furthermore, it is shown that both of these two types of texts demonstrate oral register patterns, although the informational features are more salient in translated Chinese texts than those in original Chinese texts. (2) Dimension 3 and Dimension 1 are the two strongest predictors in both translated Chinese and original Chinese texts, as is evident from figure 1. It shows translated Chinese texts and original Chinese texts demonstrate the same trend in these two dimensions at least. The register features in translated Chinese and those in original Chinese across the three sampling periods are correlated, but changed across time. English source shining-through effect in the three sampling periods has increased, as is evidenced by the number of linguistic features from the first period to the third period. (3) It is quantitatively proved that some individual features in each of the following dimensions as informational production, narrative concern, situational reference, overt expression of argument and abstract information have had an impact on the corresponding dimension in the original Chinese texts across time. However, the extent varied among the three periods.

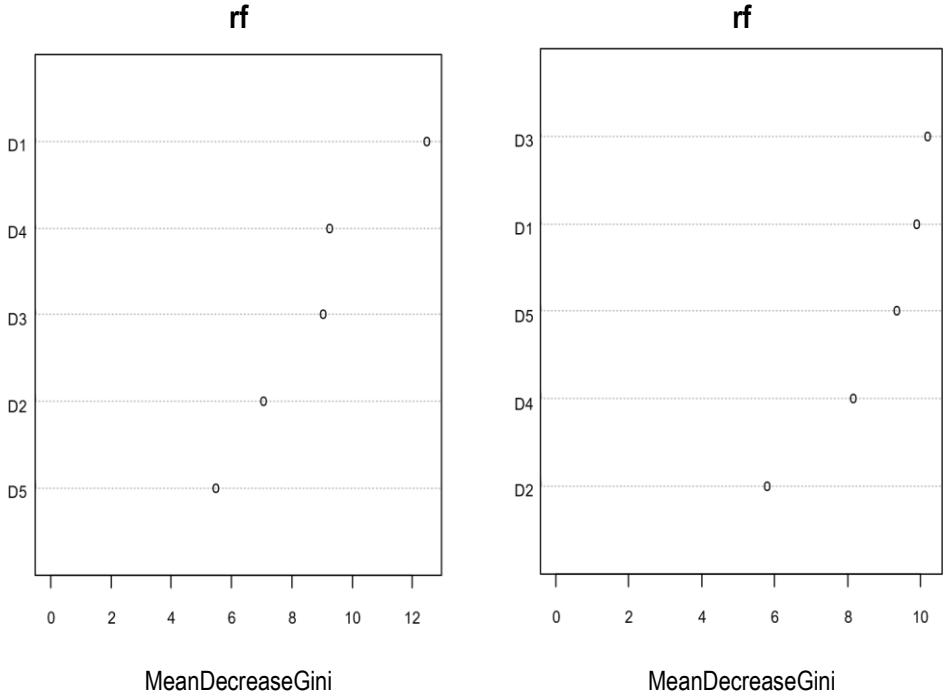


Figure 1 Variable importance plot: The relative importance of dimension scores in predicting the period of translated Chinese texts (on the left) and original Chinese texts (on the right)

Concludingly, it is argued that the register features in translated Chinese, which is the locus of language contact, on the one hand, is interfered by source language shining-through effect, and do exert influence on change of communicative functions in original Chinese on the other hand. The findings further suggest that in the early stage, the informational linguistic features in translated Chinese were normally more susceptible to change and had an effect on original Chinese texts, and then the register features in translated Chinese texts and original Chinese texts became divergent in the use of them across time. The more advanced the translated texts move along the code-copying process, the more they resemble original Chinese texts in the use of informal linguistic features in written registers due to the strength and length of language contact.

References

Amouzadeh, Mohammad and Juliane House. 2010. Translation as a Language Contact Phenomenon: The Case of English and Persian Passives. *Languages in Contrast* 10 (1): 54-75.

Baumgarten, Nicole and Demet Ozcetin. 2008. Linguistic Variation through Language Contact in Translation. In Peter Siemund and Noemi Kintana (eds). *Language Contact and Contact Languages.* Amsterdam: John Benjamins, 293–316.

Becher, Viktor. 2009. The Decline of Damit in English–German Translations. A Diachronic Perspective on Source Language Interference. *SKASE Journal of Translation and Interpretin* 4 (1): 2–24.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics* 5(1). 107-135.

Biber, D. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register vari-ation. *Languages in Contrast* 14(1): 7–34.

Bisiada, Mario. 2013. From Hypotaxis to Parataxis: An investigation of English-German Syntactic Convergence in Translation. PhD Dissertation. UK: Manchester University.

Dai, Guangrong. 2016. *Hybridity in Translated Chinese: A Corpus Analytical Framework*. Singapore: Springer.

Delaere, Isabelle, and Gert de Sutter. 2013. Applying a Multidimensional, Register-Sensitive Approach to Visualize Normalization in Translated and Non-Translated Dutch. In Marie-Aude Lefer, and Svetlana Vogeleer(eds.). *Interference and Normalization in Genre-Controlled Multilingual Corpora.* Amsterdam: Benjamins, 43–60.

Gert De Sutter & Haidee Kruger. 2018. Disentangling the motivations underlying syntactic explicitation in contact varieties. A MUPDR analysis of that vs. zero complementation. In Sylviane Granger (eds.). *Using Corpora in Contrastive and Translation Studies Conference*:55-57.

Gries, St.Th. and M. Hilpert. 2010. Modeling diachronic change in the third person singular: a multi-factorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14 (3). 293–320.

Hansen-Schirra, Silvia. 2011.Between Normalization and Shining-through: Specific Properties of English–German Translations and their Influence on the Target Language. In Svenja Kranich, Viktor Becher, Steffen Hoder and Juliane House (eds). *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*. Amsterdam: John Benjamins, 135–162.

House, Juliane. 2016. *Translation as Communication Across Languages and Cultures.* New York: Routledge.

House, Juliane. 2014. *Translation a multidisciplinary approach*. Basingstoke: Palgrave Macmillan.

Hu, Kaibao. 2005. *On the Impact of the Historical Text of English–Chinese Dictionary on the Modernization of the Chinese Language*. Foreign Languages and Their Teaching 37 (3): 57–60.

Kranich, Svenja, Viktor Becher and Steffen Hoder. 2011. A Tentative Typology of Translationinduced Language Change. In Kranich, Svenja, Viktor Becher, Steffen Hoderand Juliane House(eds.). *Multilingual Discourse Production.* Amsterdam: John Benjamins Publishing Company.

Kranich, Svenja, Juliane House and Viktor Becher. 2012. Changing Conventions in English–German Translations of Popular Scientific Texts. In Kurt Braunmüller and Christoph Gabriel (eds.). *Multilingual Individuals and Multilingual Societies*. Amsterdam: John Benjamins, 315–334.

Kruger, Haidee, and Bertus van Rooy. 2012. Register and the Features of Translated Language. *Across Languages and Cultures* 13: 33–65.

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

Kruger, Haidee. & B. van Rooy. 2018. Register variation in written contact varieties of English: a multidi-mensional analysis. *English World-Wide* 39 (2): 214–242.

Kruger, Haideee &Adam Smith. 2018. Colloquialization versus Densification in Australian: A Multidimensional Analysis of the Australian Diachronic Hansard Corpus. (ADHC). *Australian Journal of Linguistics* 38(3): 293-328.

Malamatidou, Sofia. 2016. Understanding translation as a site of language contact: the potential of the code-copying framework as a descriptive mechanism in translation studies. *Target* 28 (3), 399--423.

Redelinghuys, Karien. 2016. Levelling-Out and Register Variation in the Translations of Experienced and Inexperienced Translators: A Corpus-Based Study. *Stellenbosch Papers in Linguistics* 45: 189–220.

Reppen, R., S. Fitzmaurice & D. Biber. 2002. *Using Corpora to Explore Linguistic Variation.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Wang, Kefei. 2002. Influence of Modern Translation on Chinese. *Foreign Language Teaching and Research 34* (6): 458–463.

Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28(4):421–450.

A Bilingual Parallel Corpus for the Analysis of Students' Approach to Specialised Texts Translation

Juan-Pedro Rica-Peromingo[1], Arsenio Andrades-Moreno[2], Ángela Sáenz-Herrero[3], Sara Martínez-Portillo[1]
Universidad Complutense de Madrid[1], Universidad Politécnica de Madrid[2],
Universidad Europea Miguel de Cervantes[3]
juanpe@ucm.es, a.andrades@upm.es, asaenz@uemc.es, samart01@ucm.es

Introduction

Using corpus linguistics as a research methodology for Translation Studies has had a huge impact on translation theory and practice. It is a major tool to identify errors and analyse translation strategies in a bilingual corpus (Johansson 2007; Rica et al. 2014; Rica and Braga 2015; Maroto and Andrades 2019; Granger and Lefer, 2020; Rica *in press* 2021) as well as contributing to the process of learning languages and acquiring translation skills and abilities.

The Centre for English Corpus Linguistics of the University of Louvain, the MUST corpus (MUltilingual Student Translation Corpus) is an international project which brings together partners from a wide range of European and worldwide universities and connects Learner Corpus Research (LCR) and Translation Studies (TS). It aims to build a corpus of translations carried out by students, including both direct (L2>L1) and indirect (L1>L2) translations, from a great variety of text types, genres, and registers in a wide variety of languages. UCMA, the Spanish team from the Complutense University in Madrid, which belongs to the MUST project (Multilingual Student Translation Corpus), has been working with a corpus of students' translations for corpus linguistics and translation studies analysis.

Objectives

This paper focuses on the work carried out by the Spanish team from the Complutense University (UCMA), which is part of the MUST project and it describes the specific features of the corpus built by its members. All the texts used by UCMA are either direct or indirect translations between English and Spanish. Students' profiles comprise translation trainees, foreign language students with a major in English, MA students, all of them with different English levels (ranging from B1 to C1); for some of the students, this would be their first experience with translation.

With this corpus of translated texts, we have stated the following research questions:

- Do language learners and translator learners have the same difficulties when confronting these translations and texts?
- What types of texts present more difficulties for translation students?
- Do different learners present common mistakes in each type of texts?
- What kind of patterns or strategies do we observe in the solutions provided by the students when facing the same translations conundrums?

Methodology

We have analysed a number of translated texts from different genres and with different objectives: journalistic texts, legal texts and audiovisual texts. We have collected and catalogued those texts according to their nature: legal and specialized terminology, audiovisual texts and journalistic texts. It is worth mentioning that in the initial stages of the corpus compilation process, all the students, whose translations have been incorporated into the corpus, signed a consent form which enables us to use and analyse their texts to the ends described above. Additionally, we have collected all the necessary

metadata related to the students and the source texts in question in a separate database. In this database we quantify and qualify the results extracted from the said analysis according to the detailed parameters included in the metadata catalogue.

Once the texts are input into the Hypal database, the researchers need not only to align each of the translations carried out by the students for each source text but also annotate the texts to be able to derive any conclusions. The major tags used in the annotation process are: ST-TT Transfer, Language, Translation Procedures and Metatags. Each of these tags is subsequently subdivided in a wide range of smaller tags, which enables us to precisely categorise the translation errors within a detailed taxonomy.

Analysis

The preliminary results from the initial data obtained point out the kind of difficulties encountered by the students and reveal the most frequent strategies implemented by the learners according to their level of English, their translation experience and/or the text genres they face: audiovisual texts, legal texts or journalistic texts. Some of those strategies used by the Spanish students in their translations are: omission, reduction, functional equivalent, adaptation and transposition. The first analysis of the results show that students use different strategies according to the kind of texts they are translating. They tend to use more the omission and reduction technics in audiovisual texts and the equivalent, adaptation and transposition strategies in legal or journalistic texts. We will provide examples and verify if this tendency is maintained through a great number of texts in order to see its main causes.

We have also found common errors in the graduate and postgraduate university students' translations: following Nord's taxonomy on errors (1997: 75), we have identified transfer errors, lexical errors, grammatical errors, text-specific translation errors and cultural-related errors.

Conclusions

Translation trainees and language learners may differ in their translation methods as they are more aware of the text typologies and the possible strategies they could use in their adaptations. Also, students may reveal greater creative skills when it comes to taking certain translation risks.

Using comparative corpora will help us to gather evidence-based information on common errors and strategies employed by learners coming from different backgrounds and will provide instances of several case studies based on English to Spanish student translations in the specialized fields mentioned before. This study will also enable researchers and teachers to bring better student-oriented approaches to translation teaching methods by building more detailed taxonomies regarding errors and translations strategies.

Since all the translations provided up to this date are first instance translations of the different source texts, a further and more ambitious study, in the future, could focus on the analysis of revised translations once the errors are pointed out and studied; as it would be of great interest for the linguistic community to observe to what extent texts can be improved after being subjected to the detailed initial error analysis suggested in this study.

References

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

Johansson, S. (2007). *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Maroto N. and Andrades, A. (2019). "Corpus multilingües para la investigación y la enseñanza de la traducción: el proyecto MUST". *E-Aesla*, 5, 401-410.

Nord, C. (1997). Translating as a Purposeful Activity. Functionalist Approaches Explained. Manchester: St. Jerome.

Obrusnik, A. (2014). “Hypal: A User-Friendly Tool for Automatic Parallel Text Alignment and Error Tagging”, Eleventh International Conference Teaching and Language Corpora, Lancaster, 20th-23rd July 2014, 67-69.

Rica, J.P. (*in press* 2021). *Corpus Studies and Audiovisual Translation: Subtitling.* Series: New Trends in Translation Studies (Edited by J. Díaz Cintas). Frankfurt: Peter Lang.

Rica, J.P. and Braga, J. (2015). *Herramientas y técnicas para la traducción inglésespañol: los textos literarios*. Madrid: Escolar y Mayo.

Rica, J.P., Albarrán, R. and García, B. (2014). "New approaches to audiovisual translation: the usefulness of corpus-based studies for the teaching of dubbing and subtitling". In E. Bárcena, T. Read & J. Arús (eds.), *Languages for Specific Purposes in the Digital Area*. Berlin: Springer-Verlag, 303-322.

BrICh: a new Brazilian Portuguese-Chinese Parallel Corpus

Andressa Rodrigues Gomide[1], Tanjun Liu[2], Frederico Amorim Cavalcante[3]
Instituto Internacional da Língua Portuguesa[1], Hong Kong Baptist University[2], Universidade Federal de Minas Gerais[3]
gomide.andressa@gmail.com, liutanjun@hotmail.com, fredericoa4@gmail.com

Brazil and China are countries with increasingly stronger connections, but most of the communication between them takes place in English. Using parallel corpora for Chinese and Brazilian Portuguese (BP) would certainly diminish the demand for the use of English. However, most of the few such corpora currently available are composed of varieties of Portuguese other than BP (e.g., Xing et al. 2016). Moreover, the data in those corpora is, more often than not, compiled through machine translation (e.g., Lison & Tiedemann 2016), which compromises the accuracy of the translated data. Our research aims to fill this gap by creating a bidirectional parallel corpus of BP and Chinese featuring different text genres (e.g., news, manuals, novels and subtitles) so as to inform language learning, translator education, and to help explore and assess corpus linguistic methodologies for this language combination. Retrieving good quality translations for this language pair is not an easy task. So we decided to first run a pilot corpus compilation to acquire a better sense of (i) the types of data available, (ii) the performance of processing tools (e.g., aligners, POS taggers) on this data, and (iii) the value of the collected data for linguistic analysis and application. In this initial phase, we collected data in PB and Chinese from sources as different as Confucius Institute magazines, TEDx talks and Netflix shows subtitles, Jehovah's Witnesses publications and websites with human-translated content. The texts and metadata (e.g., genre, year of publication, author and copyright status) were stored using Timbila (Gomide forthcoming), a system for text management. Both the Chinese and BP parts of the corpus were annotated with part-of-speech and lemma information using the TreeTagger (Schmid 1994). They were then aligned using the LF-aligner (Farkas 2018) and prepared for the IMS Open Corpus Workbench (CWB; Christ 1994), which consists of a collection of tools for managing and querying large corpora. The first version of the corpus was uploaded to CQPweb (Hardie 2012), a web-based interface for CWB. CQPweb allows, among other things, an easy visualization of aligned concordance lines and metadata as well as the creation of subcorpora. During the presentation, we will outline the process of data collection and how the main challenges were addressed. In addition, we will provide a demonstration of the corpus and discuss ways in which it could be used for translation and pedagogical purposes. We will also put forth useful suggestions concerning the creation of parallel corpora of Chinese and other romance languages, such as Italian and Spanish.

References

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *arXiv preprint cmp-lg/9408005*.
Farkas, A. (2018). LF Aligner [Computer Software]. Available from https://sourceforge.net/projects/aligner/.
Gomide, A. R. (Forthcoming). *Timbila: A user-friendly web tool for corpus compilation*.
Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, *17*(3), 380-409.
Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portoroz, Slovenia, May 23-28, 2016*
Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
Xing, J., Wong, D. F., Chao, L. S., Leal, A. L. V., Schmaltz, M., & Lu, C. (2016, July). Syntaxtree aligner: A web-based parallel tree alignment toolkit. In *2016 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* (pp. 37-42). IEEE.

Pronouns in Translational Lithuanian

Jurgita Vaičenonienė, Jolanta Kovalevskaitė
Centre of Computational Linguistics, Vytautas Magnus University
jurgita.vaicenoniene@vdu.lt, jolanta.kovalevskaite@vdu.lt

Lithuanian is a synthetic language with rich inflection, high morphological ambiguity, and certain archaic features no longer preserved in other Indo-European languages. Widespread software and methods developed for the analysis of English or other analytical languages often cannot be directly applied to process Lithuanian texts. As a low-resource language, it has a continuous need for the creation of new language resources, their analysis tools, and adaptation of existing software and methodological approaches. This gave an impetus to develop language resources tailored for the needs of Translation Studies such as ORVELIT[13], a comparable corpus of original and translated Lithuanian fiction and popular science. In previous research, we described the composition, balance, and representativeness of the corpus (Vaičenonienė et al., 2017) and provided a quantitative analysis of the lexical and morphological features of translational Lithuanian (Vaičenonienė and Kovalevskaitė, 2019). We observed that in fiction and popular science translations from English, there are significantly more pronouns than in the original Lithuanian texts. However, it remains unclear which specific types of pronouns are over- or under-used and in which particular contexts. This paper aims to compare the distribution of specific semantic subclasses of pronouns in original and translated Lithuanian, specifically focusing on the use of dual pronouns.

To extract pronouns from the morphologically annotated version of the ORVELIT v2 corpus, we used the GNU Grep (general regular expression pattern matcher) tool. The pronouns were further manually classified into the following semantic subclasses (based on Ambrazas et al. 2006):

Table 1 Pronouns in the ORVELIT corpus[14]

Semantic subclass		**Original popular science subcorpus**	**Translated popular science subcorpus**	**Original fiction subcorpus**	**Translated fiction subcorpus**
Personal	Proper	25653	45296	62240	80628
	Possessive	262	236	418	308
	Reflexive	4484	7322	5259	4711
Demonstrative		18420	23025	19912	20580
Indefinite	Proper	3192	4904	4448	5296
	Generalizing	6498	8937	11414	10785
Interrogative-Reflexive		10666	14098	13611	13499
Total:		**78491**	**114328**	**126277**	**143631**

As is seen in Table 1, pronoun frequencies and types vary in the two registers with a tendency to be over- rather than under-used in translation. This particularly applies to proper personal and proper indefinite pronouns in both fiction and popular science translations and reflexive personal, demonstrative, generalizing indefinite and interrogative reflexive pronouns in translated popular science. A detailed analysis of the selected semantic subclasses will allow us to detect deviant pronoun types which are going

[13] https://clarin.vdu.lt/xmlui/handle/20.500.11821/40

[14] As each subcorpus of original and translated fiction and popular science consists of approx. 1 mln. words, normalized frequencies were not calculated.

to be further analysed by looking into their contextual environment in comparable and parallel concordances. Parallel Corpus[15] (2024999 tokens for English-to-Lithuanian translations) will be used to retrieve original English sentences for the subcorpus of translated fiction.

A preliminary analysis of pronoun types in each semantic subclass also shows variation. For example, some lemmas, relatively infrequent in original Lithuanian, are even less common or absent in translations (e.g., demonstratives *anas* 'that one'; *anoks* 'like that one/ of that type'; a generalizing indefinite *tūlas* 'more than one'). Another trend is that translations exhibit visibly higher frequencies of personal proper pronouns such as *aš* 'I', *jis* 'he', *tu/jūs* 'you' and the first person dual *mudu* 'two of us'. This overuse of the dual form is particularly interesting as, considering previous research, unique target language items should occur less frequently in translations (Trikkonen-Condit 2004). Therefore, we further calculated the frequencies of all dual pronouns in the corpus:

Table 2 Frequencies (tokens) of dual pronouns in the ORVELIT corpus

Original popular science subcorpus	Translated popular science subcorpus	Original fiction subcorpus	Translated fiction subcorpus
34	48	426	724

Differently from English, some Lithuanian nominal and adjectival pronouns have a dual number (e.g., *mudu* 'two of us', *judu* 'two of you', *katriedu* 'two of which'), seen as a subtype of plural (Ambrazas et al. 2006: 184). Functionally, duality is chosen over the plural form to put emphasis on the two specific participants, or to show the close relationship between the two referents (Roduner and Čižik, 2006: 75-78). The use of dual pronouns is an optional rather than an obligatory grammatical choice, it is more of a semantic or stylistic rather than grammatical nature and depends on how the speaker/ writer understands or wants to emphasize the proximity/distance relations of the referents (ibid.). As English does not have the grammatical category of duality, Lithuanian dual pronouns can be seen as unique items. Trikkonen-Condit (2004: 177-178) maintains that phenomena existing in the grammatical, lexical, or other patterning of the target language, but absent or manifested differently in the source language "do not suggest themselves as translation equivalents as there is no obvious linguistic stimulus for them in the source text". As a result, translations are more likely to have lower frequencies of these unique items in comparison to the original texts of the target language (ibid.). However, as is seen in Table 2, in the ORVELIT v2 corpus, dual pronouns occur significantly more frequently in translated fiction.

Further on, we will briefly overview the Lithuanian pronominal system by highlighting some features which are realized differently in English, present the factors relevant for the analysis of English to Lithuanian translations, and provide a detailed discussion of the results. Insights in relation to prior research on different types of source language interference (Eskola 2004, Trikkonen-Condit 2004, Teich 2003, Toury 1995) will also be given.

References

Ambrazas, V., E. Geniušienė, A. Girdenis, N. Sližienė, D. Tekorienė, A. Valeckienė, & Valiulytė, E. (2006). *Lithuanian Grammar*. Vilnius: Baltos Lankos.

Eskola, S. (2004). Untypical Frequencies in Translated Language: A Corpus-based Study on Literary Corpus of Translated and Non-Translated Finnish. In A. Mauraneen, & P. Kujamäki (eds.), *Translation Universals. Do they Exist?* Amsterdam & Philadelphia: John Benjamins, 83–99.

Grep for Windows. (2009). http://gnuwin32.sourceforge.net/packages/grep.htm (Accessed 10 01 2020)

Roduner, M., & Čižik-Prokaševa, V. (2006). Skaičiaus kategorija. In A. Holvoet and R. Mikulskas (eds.), *Gramatinių funkcijų tyrimai.*, 67–100.

[15] https://klc.vdu.lt/lygiagretus-tekstynas/

Teich, E. (2003). *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.
Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam & Philadelphia: John Benjamins.
Trikkonen-Condit, S. (2004). Unique Items — Over- or Under-represented in Translated Language? In A. Mauraneen, & P. Kujamäki (eds.), *Translation Universals. Do they Exist?* Amsterdam & Philadelphia: John Benjamins, 177–184.
Vaičenonienė, J., & Kovalevskaitė, J. (2019). Leksinės ir morfologinės vertimų kalbos ypatybės (Lexical and Morphological Features of Translational Lithuanian). In *Darnioji daugiakalbystė/ Sustainable Multilingualism 14*, 208–235.
Vaičenonienė, J., J. Kovalevskaitė, & Ringailienė, T. (2017). Tekstynais paremti vertimų kalbos tyrimai ir šaltiniai (Corpus-Based Translation Studies: Research and Resources). In *Kalbų studijos/ Studies about Languages 30*, 42–55.

Translation-induced language variation and change: A case study of the impact of English-Chinese translations on delexicalized verb *zuo* (do) from a diachronic perspective

Qiurong Zhao, Xuee Xie
University of Science and Technology Beijing, China
qiurong.zhao@ustb.edu.cn, xuee_xie@163.com

Translation-induced language contact (Kranich 2014) has witnessed various degrees of transfer of linguistic features (House 2006; Steiner 2008; Kranich et al. 2012), while less empirical research has been performed on translation-induced language variation and change across distant language pairs, for instance English-Chinese translation.

The present paper contributes to triangulating corpus-based study of language contact through translation by constructing diachronic multiple corpora, which consists of: (a) the 6.97-million-word comparable corpora (1949-2019), including translated Chinese texts and non-translated Chinese texts. It covers three genres fictions, social sciences texts and scientific texts by three different time periods: 1949-1959, 1979-1989 and 2009-2019; (b) a 1-million-word parallel corpus (2009-2019) of English scientific texts and their Chinese translations.

Delexicalized verbs in modern Chinese language have become more and more frequent, partly due to the Europeanization of Chinese (Yu 2000) or language contact (Wei 2007). Among these verbs, *zuo* (做 "do") is especially frequently used. The present study aims to explore the following questions:

(1) What are the differences of delexicalized verb *zuo* between translated Chinese and non-translated Chinese over time?
(2) To what degree translation has influenced the usage of delexicalized verb *zuo* ? And what is the interaction mode of shining-through (Teich 2003) and normalization (Baker 1996; Hansen-Schirra & Steiner 2012) in translating *zuo*?

This study proceeds as follows: Firstly, diachronic frequency of delexicalized verb *zuo* are examined in translated and non-translated texts. Secondly, delexicalized verb *zuo* are investigated in three different genres to determine the influence of genre in its development. Thirdly, the parallel corpus is applied to investigate what expressions in English source texts trigger the use of Chinse *zuo*. And multifactorial analysis based on the conditional inference tree, the random forest and cluster analysis have been performed to assess the potential impact of language-independent and language-dependent factors on the translation of *zuo* and to what extent translation is an important factor in the variation and change of *zuo* in this aspect.

The findings show that
(1) Both translated Chinese and non-translated native Chinese have seen the increased frequency of delexicalized verb *zuo*, in particular, in translated Chinese. In three genres, *zuo* has significantly increased and it is not sensitive to genres.
(2) About 51% of *zuo* are translated from English delexicalized verb structures and their collocations are equivalent to the source language, which may be a demonstration of the source text shining-through. Surprisingly, 49% *zuo* are not triggered by English source structures. They are added or explicated by translators due to ensure fluency, clarity or avoid repetition in Chinese, which may be an evidence of adaptation to the norms in Chinese language and risk aversion in translation. Translation can be a factor in the diocronic change of *zuo*, especially in the given favorable social context. And other language-

dependent and language-independent factors also play a role in the variation and change of delexicalized verb *zuo*.

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (ed.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam: John Benjamins. 175-186.

Hansen-Schirra, S. & Steiner, E. (2012). Towards a typology of translation properties. In S. Hansen-Schirra, S. Neumann, & E. Steiner (eds.) *Cross-linguistic corpora for the study of translations: Insights from language pair English-German*. Berlin: Walter de Gruyter. 255-279.

House, J. (2006). Covert Translation, Language Contact, Variation and Change. *SYNAPS* 19, 25-47.

Kranich, S., J. House, and V. Becher. (2012). Changing Conventions in English-German Translations of Popular Scientific Texts. In K. Braunmüller & C. Gabriel (eds.) *Multilingual Individuals and Multilingual Societies*. Amsterdam: John Benjamins Publishing Company. 315-334.

Kranich S. (2014). Translations as a locus of language contact. In House J. (ed) *Translation: A multidisciplinary approach. Palgrave advances in language and linguistics*. Palgrave Macmillan, London. 96-115.

Steiner, E. (2008). "Empirical Studies of Translations as a Mode of Language Contact". In P. Siemund & N. Kintana (eds.) *Language Contact and Contact Languages*. Amsterdam: John Benjamins Publishing Company. 317-345.

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

Wei, N. X. (2007). Shared Meaning and Delexicalization. *Journal of PLA University of Foreign Languages* 5. 17-24. [In Chinese].

Yu, G. Z. (2000). *Yu Guangzhong's Papers on Translation*. Beijing: China Translation Corporation. 152-154. [In Chinese].

Contrasting corpora to identify phraseological suggestions to enhance L2 English research writing

Gustavo Zomer, Ana Frankenberg-Garcia
University of Surrey
g.zomer@surrey.ac.uk, a.frankenberg-garcia@surrey.ac.uk

Research writing is a challenge for most people, including experienced academics. Given that English has become a worldwide lingua franca of science communication, research writing can be especially challenging for scholars and scientists whose habitual working language is not English (L2 English researchers) (Schuster, Levkowitz, & Oliveira Junior, 2014; Politzer-Ahles, Holliday, Girolamo, Spychalska, & Berkson, 2016). Although even speakers whose first language (L1) is English may struggle with academic English (Hyland, 2006; Kosem, 2010; FrankenbergGarcia, 2018), L2 English researchers are likely to require more effort to write about their work, since they receive less incidental exposure to the target language.

A major hurdle in the way of L2 English researchers when writing for publication in highranking international journals is academic English phraseology. Language is one of the assessment parameters in the peer-review process, and it has been argued that, regardless of contents, unconscious bias makes reviewers and editors more likely to accept papers that read well than ones with "poor" English (Politzer-Ahles et al., 2016; Flowerdew, 2019). Perhaps more importantly, the use of conventional phraseology facilitates language processing (Conklin & Schmitt, 2012). As such, the readability of complex texts like research papers can be compromised by a less idiomatic use of academic English, which could in turn limit L2 English researchers' chances of publication, and, if published, their chances of citation.

There are nevertheless numerous tools and resources to help enhance academic English phraseology, including textbooks, dictionaries, corpora and writing assistants. However, not many writing aids have been customized to address the specific needs of L2 researchers from a given L1 background. Moreover, most tools aimed at addressing L2 difficulties have focused on errors (Chang, Chang, Chen, & Liou, 2008; Futagi, Deane, Chodorow, & Tetreault, 2008; Gao, 2013; Huang & Tsao, 2019), and draw mainly on small, error-annotated corpora of intermediate learners of English (Chang et al., 2008; Futagi et al., 2008), typically from undergraduate essays (Gao, 2013) or English certificate exams (Huang & Tsao, 2019). There does not seem to be much in the way of support for researchers who have a good command of basic English but still experience phraseological limitations related to their specific L1 background.

This study aims to help Brazilian researchers improve their phraseological repertoire when writing for publication in English. Unlike studies that draw on error-annotated corpora of student writing, our goal is to automatically compare the phraseological profile of mature Brazilian researchers writing in English with that of researchers that publish internationally, and provide target English solutions that can assist the former. In doing so, the study addresses the following research questions:

(i) What are the main phraseological differences in journal articles published locally by Brazilian researchers when compared with a reference corpus of expert academic English?
(ii) Can L1 Portuguese academic phraseology explain some of the discrepancies above?
(iii) How to automatically provide target academic English phraseological solutions to help Brazilian researchers writing for publication in English?

Three corpora are used in the study (Figure 1). The focus corpus, BrACE v.2, is a much larger, 35M word version of the 1M word BrACE corpus used in Tavares Pinto, Rees & Frankenberg-Garcia (2021). It was

compiled using a balanced sample of Brazilian journal articles in seven broad subject areas downloaded from *Scientific Electronic Library Online* (SciELO). SciELO provides access to a good number of Brazilian journals whose language is English or which include articles written in English. Unlike many of the sources used to investigate L2 English phraseology, the texts mined for BrACE are authentic full-length articles authored by active L2 English researchers reporting on real studies in a wide range of areas.

The reference corpus is the Expert Academic Corpus of English (ExpACE), which was built specifically for this study. It consists of 35M words from more than 5000 highly cited papers published in high-impact journals in eight subject areas. The third corpus is CoPEP, a 40M word corpus of academic Portuguese sourced from published journal articles from SciELO (Kuhn, 2017).

Figure 1. Corpora used in the study

To identify phraseological contrasts between publications by L2 English Brazilian researchers and expert academic English (RQ1), we follow Granger's (Granger, 1996) Contrastive Interlanguage Analysis. We began by extracting the top 2-4 grams in BrACE and ExpACE with a normalized frequency of over ten per million, selecting only n-grams occurring in more than three disciplines in each corpus, and removing proper nouns and non-phrases. Next, we contrasted BrACE with ExpACE using a smoothed-frequency ratio (Kilgarriff, 2009), and selected the n-grams that occurred twice as often in the former to look for evidence of overuse. Then we performed a similar procedure to identify underused n-grams by employing BrACE as reference and ExpACE as the focus corpus.

To identify the possible sources of n-gram overuse that could be related to L1 (RQ2), we machine-translated them into Brazilian Portuguese using DeepL, and selected the translation that occurred most frequently in the CoPEP corpus. Finally, we machine-translated each Brazilian Portuguese phrase back to English, which resulted in alternative lexical suggestions for overused English phrases (RQ3).

Preliminary results suggest that many phraseological contrasts observed can be traced back to academic Portuguese. For example, *according to* is almost four times more frequent in BrACE than in the expert corpus, and its direct Portuguese translation *de acordo com* is the most frequent equivalent in academic Portuguese. By machine-translating it back to English, it is possible to automatically suggest phrase replacements, like replacing *according to these results* with *in line with these results*.

Overused and underused phraseology is not always obvious to detect and address. This study fills this gap by developing a contrastive approach for automatically identifying typical issues and offering alternative suggestions without using error-annotated corpora. Our methodology can be extended for researchers with different L1 backgrounds by changing the L2 English corpus and L1 academic corpus.

References

Chang, Y. C., Chang, J. S., Chen, H. J., & Liou, H. C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3), 283–299.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32, 45–61.

Flowerdew, J. (2019). The linguistic disadvantage of scholars who write in English as an additional language: Myth or reality. *Language Teaching*, 52(2), 249-260.

Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes*, 35, 93–104.

Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21(4), 353–367.

Gao, Z. M. (2013). Automatic identification of English collocation errors based on dependency relations. *27th Pacific Asia Conference on Language, Information, and Computation, PACLIC 27*, 550–555.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In *Languages in contrast. Text-based cross-linguistic studies*. Lund: Lund University Press.

Huang, P. Y., & Tsao, N. L. (2019). Using collocation clusters to detect and correct English L2 learners' collocation errors. *Computer Assisted Language Learning*, 0(0), 1–27.

Hyland, K. (2006). *English for Academic Purposes*. New York: Routledge.

Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK.

Kosem, I. (2010). *Designing a model for a corpus-driven dictionary of Academic English*. Birmingham: Aston University.

Kuhn, T. (2017). *A design proposal of an online corpus-driven dictionary of Portuguese for university students*. University of Lisbon.

Politzer-Ahles, S., Holliday, J. J., Girolamo, T., Spychalska, M., & Berkson, K. H. (2016). Is linguistic injustice a myth? A response to Hyland (2016). *Journal of Second Language Writing*, 34, 3–8.

Schuster, E., Levkowitz, H., & Oliveira Junior, O. N. de. (2014). *Writing scientific papers in English successfully: your complete roadmap*. São Carlos, Brazil: Compacta.

Tavares Pinto, P., Rees, G. P., & Frankenberg-Garcia, A. (2021.). Identifying collocation issues in English L2 research article writing. In M. Charles & A. Frankenberg-Garcia (Eds.), *Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis*. London: Routledge, 147-170.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF
INTERPRETING AND TRANSLATION