CECL Papers 1

Book of Abstracts

Using Corpora in Contrastive and Translation Studies Conference (5th edition)

Sylviane Granger, Marie-Aude Lefer and Laura Aguiar de Souza Penha Marion (eds)

Centre for English Corpus Linguistics

UCLouvain

Louvain-la-Neuve, 12-14 September, 2018

Book of Abstracts

Using Corpora in Contrastive and Translation Studies Conference (5th edition)

Louvain-la-Neuve, 12-14 September, 2018

Sylviane Granger, Marie-Aude Lefer and Laura Aguiar de Souza Penha Marion (eds)

CECL Papers 1. Louvain-la-Neuve: Université catholique de Louvain
2018

Organizing committee

Conference Chairs
Sylviane Granger (Université catholique de Louvain)
Marie-Aude Lefer (Université catholique de Louvain)

Laura Aguiar de Souza Penha Marion (Université catholique de Louvain)
Maïté Dupont (Université catholique de Louvain)
Gaëtanelle Gilquin (Université catholique de Louvain)
Christine Michaux (Université de Mons)
Magali Paquot (Université catholique de Louvain)

Scientific committee

Silvia Bernardini (University of Bologna)
Łucja Biel (University of Warsaw)
Bert Cappelle (Université de Lille 3)
Andrew Chesterman (University of Helsinki)
Lucie Chlumská (Charles University)
Hélène Chuquet (Université de Poitiers)
Jean-Pierre Colson (Université catholique de Louvain)
Gloria Corpas Pastor (University of Malaga)
Barbara De Cock (Université catholique de Louvain)
Sabine De Knop (Université Saint-Louis – Bruxelles)
María de los Ángeles Gómez González (Universidad de Santiago de Compostela)
Gert De Sutter (Ghent University)
Bart Defrancq (Ghent University)
Liesbeth Degand (Université catholique de Louvain)
Isabelle Delaere (KU Leuven)
Ilse Depraetere (Université de Lille 3)
Pamela Faber (University of Granada)
Adriano Ferraresi (University of Bologna)
Kjersti Fløttum (University of Bergen)
Thierry Fontenelle (Centre de traduction des organes de l'Union européenne)
Ana Frankenberg-Garcia (University of Surrey)
Federico Gaspari (University for Foreigners "Dante Alighieri" of Reggio Calabria)
Volker Gast (University of Jena)
Gaëtanelle Gilquin (Université catholique de Louvain)
Patrick Goethals (Ghent University)
Sandra Halverson (Western Norway University of Applied Sciences)
Silvia Hansen-Schirra (Johannes Gutenberg University of Mainz)
Andrew Hardie (Lancaster University)
Hilde Hasselgård (University of Oslo)
Philippe Hiligsmann (Université catholique de Louvain)
Véronique Hoste (Ghent University)
Juliane House (University of Hamburg)
Matthias Hüning (Free University of Berlin)
Marta Kajzer-Wietrzny (Adam Mickiewicz University)
Dorothy Kenny (Dublin City University)
Haidee Kruger (Macquarie University)
Natalie Kübler (Université Paris-Diderot)
Kerstin Kunz (University of Heidelberg)
Béatrice Lamiroy (KU Leuven)

Ekaterina Lapshinova-Koltunski (Saarland University)
Sara Laviosa (University of Bari Aldo Moro)
Torsten Leuschner (Ghent University)
Diana Lewis (Université de Provence Aix-Marseille I)
Defeng Li (University of Macau)
Rudy Loock (Université de Lille 3)
Markéta Malá (Charles University)
Josep Marco (Jaume I University)
Juana Isabel Marín Arrese (Universidad Complutense de Madrid)
Adriana Mezeg (University of Ljubljana)
Christine Michaux (Université de Mons)
Ruslan Mitkov (University of Wolverhampton)
Stella Neumann (RWTH Aachen University)
Raluca Nita (Université de Poitiers)
Signe Oksefjell Ebeling (University of Oslo)
Magali Paquot (Université catholique de Louvain)
Rosa Rabadán (University of León)
Raf Salkie (University of Brighton)
Erich Steiner (Saarland University)
Frieda Steurs (KU Leuven)
Elke Teich (Saarland University)
Aurelija Usonienė (Vilnius University)
Kristel Van Goethem (Université catholique de Louvain)
Sonia Vandepitte (Ghent University)
Gudrun Vanderbauwhede (Université de Mons)
Åke Viberg (Uppsala University)
Svetlana Vogeleer (Université catholique de Louvain / Université Saint-Louis – Bruxelles)
Geoffrey Williams (Université de Bretagne-Sud / Université Grenoble Alpes)
Federico Zanettin (University of Perugia)
Sandrine Zufferey (Université de Berne)

Acknowledgments

We would like to thank our sponsors for their support of the conference.

- Fonds National de la Recherche Scientifique (F.R.S.-FNRS)
- Institute for Language and Communication/Linguistics Unit (UCLouvain)
- Faculty of Philosophy, Arts and Letters/Linguistics programme (UCLouvain)
- Common Language Resources and Technology Infrastructure (CLARIN)
- Learner Corpus Association
- John Benjamins Publishing Company
- Bloomsbury Publishing Company
- Edinburgh University Press
- Lexical Computing

Table of contents

Keynote presentations

Papers and posters

Keynote presentations

In principio erat Verbum: A fresh look at corpora for translation and interpreting

Gloria Corpas Pastor
University of Malaga, University of Wolverhampton
gcorpas@uma.es

This paper intends to outline the state of the art of corpus-based translation and interpreting with a view to identifying new challenges and research opportunities. From ancient times, *verba* ('words') and the messages they convey have played a very important role in mediated discourse. Just equipped with the power of words, over the years translators and interpreters have practiced their work on a daily basis. Both have relied heavily on dictionaries, glossaries, term spreadsheets and the like. Later on, e-resources and language technologies became translators' best friends.

Nowadays, language technologies play a fundamental role in translators' workflows. Tech-savviness is no longer a rare asset, but the industry is already looking for new profiles, i.e. translators who are also qualified information technology experts and/or fulfill the requirements of new job profiles (e.g. post-editing). As Bowker & Corpas Pastor (2015) say: "In today's market, the use of technology by translators is no longer a luxury but a necessity if they are to meet rising market demands for the quick delivery of high-quality texts in many languages."

Translators use a wide range of electronic tools and resources (including corpora) that help them carry out various translation-related tasks, as well as CAT tools proper (translation memories, machine translation systems, localisation tools, etc.), either standalone or bundled into a tool suite. Some individual tools are more automated, more expensive and require a steeper learning curve than others. Those are determining factors that explain translators' different habits, trends and degrees of technology uptake (cf. Zaretskaya et al. 2018). Corpora and corpus management tools appear to be particularly accessible and easy to use. The advantages of using corpora in the work and training of translators have already been shown by early studies (cf. Granger et al. 2003, Zanettin et al. 2003, among others; see also Fantinuoli & Zanettin 2015 and Corpas Pastor & Seghiri 2016). Some of the principal advantages of using corpora are their reusability, modularity and flexibility. In addition, these are rather inexpensive aids that allow easy access to and management of huge quantities of information in almost no time. Corpora are being increasingly used in translation training as life-long learning resources, to analyse differential performance and naturalness in trainees' outputs (as compared with non-translated language), and to promote biculturalism. Finally, corpus-based translation studies have become a major paradigm in research methodology since the last decade (Corpas Pastor 2008): from the study of universals, translationese and process-oriented analysis, including different translation settings (Fantinuolli & Zanettin 2015), to tighter integration of CAT tools and corpora (e.g. neural/adaptive machine translation, linguistically enriched translation memories, etc.).

Interpreters, by contrast, have rarely benefited from language technologies and tools to make their work more efficient (Costa et al. 2014). Although most interpreters are unaware of interpreting technologies or are reluctant to use them (Corpas Pastor & Fern 2016), there are some tools and resources already available (Sandrelli 2015; Fantinuolli 2018). Major concerns are the loss of quality and the dehumanisation of interpreting that allegedly tend to accompany technological developments (Jourdenais & Mikkelson 2015). Nowadays, computer-assisted interpreting (CAI) tools basically encompass terminology management tools, corpora, note-taking applications and converters. However, current technological advances in interpreting differ so much from interpreters' work practice that they are perceived as irrelevant or useless. Two types of CAI tools are currently gaining ground: terminology and corpus management tools. This should not come as a surprise given the special role of specialised terminology (domain and lexical knowledge) in the preparation phase (Costa et al. 2017). The advantages of a corpus-driven approach to interpreting preparation and interpreting quality (especially as regards terminology and phraseology) have been pointed out by Aston (2015), Fantinuolli (2017) and Pérez-Pérez (2018), among others. See also the papers in the edited volume by Straniero & Falbo (2012).

Nevertheless, corpora still present many shortcomings for interpreting purposes. On the one hand, compiling interpreting corpora is a complex, challenging and time-consuming activity, especially in comparison with translation corpora: "The recording and transcription of unscripted speech events is highly labour intensive in comparison to the work involved in collecting quantities of written text for analysis" (Thompson 2005: 254). The few existing ones are not usually based on authentic interpreting, rather on parallel corpora of translations, and do not tend to contain an aligned oral component. Few collections of interpreting data are available, and those tend to be too small, too narrow in focus and not representative enough. Besides, transcription of spoken data for corpus compilation is also a time-consuming process and multimodality remains a serious problem. Multimodal corpora have been frequently treated only marginally within the field of corpus linguistics and the support for multimodal corpora is limited in most corpus management systems. These are just some of the technical challenges that corpus-based interpreting faces at the moment. Further challenges that will also shape the future of the interpreting profession are the integration of corpora with other CAI and NLP tools and the automation of interpreting solutions that will certainly follow.

References

Aston, G. (2015). Learning phraseology from speech corpora. In A. Leńko-Szymańska & A. Boulton (eds).. *Multiple Affordances of Language Corpora for Data-driven Learning*. (Studies in Corpus Linguistics 69), 63-84.

Bowker, L. & Corpas Pastor, G. (2015). Translation Technology. In R. Mitkov (ed.) *Handbook of Computational Linguistics*. Oxford: Oxford University Press. Available online from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199573691.001.0001/oxfordhb-9780199573691-e-007 [Accessed 2018-05-02].

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt: Peter Lang.

Corpas Pastor, G. & Durán Muñoz, I. (eds). (2017). *Trends in E-tools and Resources for Translators and Interpreters*. (Approaches to Translation Studies, 45). Leiden: Brill Rodopi.

Corpas Pastor, G. & Fern, L. (2016). *A survey of interpreters' needs and practices related to language technology*. Technical paper [FFI2012-38881-MINECO/TI-DT-2016-1]. University of Malaga. Available online from http://www.lexytrad.es/assets/Corpas-Fern-2016.pdf [Accessed 2018-04-03].

Corpas Pastor, G. & Seghiri Domínguez, M. (eds). (2016). *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*. (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, 106). Frankfurt: Peter Lang.

Costa, H., Corpas Pastor, G. & Durán Muñoz, I. (2014). Technology-assisted Interpreting. *Multilingual* 143(25), 27-32

Costa, H., Corpas Pastor, G. & Durán Muñoz, I. (2017). Assessing Terminology Management Systems for Interpreters. In G. Corpas Pastor & I. Durán Muñoz (eds). *Trends in e-tools and resources for translators and interpreters*. Leiden: Brill, 57-84.

Fantinuoli, C. (2017). Computer-assisted preparation in conference interpreting. *The International Journal for Translation and Interpreting Research* 9(2), 24-37.

Fantinuoli, C. (2018). Computer-assisted Interpreting: Challenges and Future Perspectives. In G. Corpas Pastor & I. Durán Muñoz (eds). *Trends in E-tools and Resources for Translators and Interpreters*. Leiden: Brill, 153-174.

Fantinuoli, C. & Zanettin, F. (eds). (2015). *New directions in corpus-based translation studies* (Translation and Multilingual Natural Language Processing 1). Berlin: Language Science Press. Available online from http://langsci-press.org/catalog/book/76 [Accessed 2018-05-02].

Granger, S., Lerot, J. & Petch-Tyson, S. (eds). (2003). *Corpus Based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi.

Jourdenais, R. & Mikkelson, H., (2015). Conclusion. In *Routledge Handbook of Interpreting*. London/New York: Routledge, 447-450.

Pérez-Pérez, P. (2018). The Use of a Corpus Management Tool for the Preparation of Interpreting Assignments: A case study. *The International Journal for Translation and Interpreting Research* 10(1), 137-151.

Sandrelli, A. (2015). Becoming an Interpreter: The role of computer technology. *MonTI. Monografías de Traducción e Interpretación* 2, 111-138.

Straniero S. & Falbo, C. (eds). (2012). *Breaking Ground in Corpus-based Interpreting Studies*. Bern: Peter Lang.

Thompson, P. (2005). Spoken Language Corpora. In W. Wyne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 59-70. Available online from http://ota.ox.ac.uk/documents/creating/dlc/ [Accessed 2018-05-02].

Zanettin, F., Bernardini, S., Stewart, D. (eds). (2003). *Corpora in Translator Education*. Manchester: St Jerome.

Zaretskaya, A., Corpas Pastor, G. & Seghiri, M. (2018). User Perspective on Translation Tools: Findings of a User Survey. In G. Corpas Pastor & I. Durán Muñoz (eds). *Trends in E-tools and Resources for Translators and Interpreters*. Leiden: Brill, 37-56.

Cognitive translation studies and the combination of data types and methods

Sandra L. Halverson
Western Norway University of Applied Sciences
Sandra.Louise.Halverson@hvl.no

Cognitive approaches to the study of translation are showing a similar development to that within Cognitive Linguistics (CL) and psycholinguistics, where multiple data types are increasingly being combined in empirical investigations. Several recent studies in Translation Studies (TS) have combined corpus data and psycholinguistic experiments or other types of observational data such as keystroke logs or eyetracking data. This development is promising in several regards: most importantly, it supports theoretical development and provides richer understanding of the data.

In this talk, I will begin by outlining the rationale for this empirical strategy from the starting point of general research methodology and from within the cognitive linguistic approach. The rationale provided for CL has been provided by such scholars as Divjak & Arppe (2013), Gilquin & Gries (2009), Heylen et al. (2008) and Tummers et al. (2005). Basically, the situation is this: cognitive theoreticians wish to explain linguistic structure and function through reference to cognitive structures and processes. Most linguistic data is, however, offline: it is the result of production processes that are not directly observable. In other words, product data do not provide unambiguous grounds for drawing inferences about cognitive structures or online processes. The question facing cognitive scholars is thus how to relate the patterns and relationships that are found in production data (for instance, a corpus) to the kinds of explanatory factors that cognitive theorists wish to evoke. A common starting point is outlined by Divjak & Arppe as follows:

> Although corpus data do not reflect the characteristics of mental grammars directly, we do consider corpus data a legitimate source of data about mental grammars. Since the results of linguistic cognitive processes, e.g. corpus data, are not independent of, or unrelated to, the linguistic knowledge that is represented in the brain, we may assume with justification that characteristics observable in language usage reflect characteristics of the mental processes and structures yielding usage, even though we do not know the exact form of these mental representations. (2013: 229-230).

The methodological tack taken by Divjak & Arppe (2013) and also proposed by Gilquin & Gries (2009) and others is to combine corpus studies with other types of observational data (e.g. elicitation data, grammaticality judgements, sorting tasks) and/or psycholinguistic tests. A central concern in this regard is the selection of the theoretical constructs used to ground the combination of methods. Divjak and Arppe's study investigated category structure, and other studies have looked at, for example structural priming (Gries 2005) or bilingual lexical relationships (Prior et al. 2011). All of these constructs are now being incorporated into cognitive TS.

In addition to the increasing use of multimethod studies, TS has also seen the more frequent use of advanced statistical techniques. In this presentation, I will include mention of developing statistical methods that support new types of corpus analysis, including multivariate statistics and the investigation of distributional characteristics (e.g. behavioral profiles), and measures of dispersion and co-occurrence relationships. The exemplary studies to illustrate this development are Vandevoorde (2016), Szymor (2017) and Schaeffer et al. (2016). In addition to illustrating the opportunities provided by the new methods, emphasis is also placed here on the constructs used to ground the studies within TS. Semantic structure, chunking, and structural priming figure here.

After presenting these two contemporary developments, I will sketch out the development of a set of theoretical constructs that I have proposed for the investigation of aggregate features of translated text (and possibly bilingual text production). The discussion will focus on the constructs themselves and how they might function in grounding multimethod studies. The three constructs, *gravitational pull*, *magnetism* and *connectivity*, were presented in Halverson (2017), though gravitational pull was proposed in an earlier paper (Halverson 2003). In the 2017 study, a range of methods were used, some of them somewhat rudimentary. In this talk, I will suggest some additional measures and data types that might lead to better testing of the hypotheses linked to these constructs. As will be demonstrated, both *gravitational pull* and *magnetism* can most economically be considered to be frequency effects, while *connectivity* may be linked to the psycholinguistic idea of 'translation ambiguity' and the 'dominant translation' (e.g. Prior et al. 2011). This discussion will also include mention of other studies and projects in TS that aim to combine corpus studies and process-based investigative methods. The best examples here are the TRICKLET project in Aachen and the CRITT Translation Process Research Database.

In closing, I will try to characterize the scientific gains that the two methodological developments have led to, and to look down the road a bit to see where TS in general, and cognitive TS in particular, might be headed as regards the development of theoretical constructs and empirical approaches. In terms of the range and sophistication of methods and approaches used, the past twenty years have witnessed an incredible rate of development, and the work done by the next generation of TS scholars is evidence that this trend is a lasting one.

References

Divjak, D. & A. Arppe. (2013). Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive linguistics* 24(2), 221-274.

Gilquin, G. & S. Gries. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1), 1-26.

Gries, S. (2005). Syntactic priming: a corpus-based approach. *Journal of Psycholinguistic Research* 34(4), 365-399.

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.

Heylen, K., J. Tummers & D. Geeraerts. (2008). Methodological issues in corpus-based cognitive linguistics. In G. Kristiansen & R. Dirven (eds). *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems.* Berlin: Mouton de Gruyter, 91-128.

Prior, A., S. Wintner, B. Macwhinney & A. Lavie. (2011). Translation ambiguity in and out of context. *Applied Psycholinguistics* 32, 93-111.

Schaeffer, M., B. Dragsted, K. Tangsgaard Hvelplund, L. W. Balling & M. Carl. (2016). Word translation entropy: Evidence of early target language activation during reading for translation. In M. Carl, S. Bangalore & M. Schaeffer (eds). *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB.* Heidelberg: Springer, 183-210.

Szymor, N. (2017). Translation: universals or cognition? A usage-based perspective. *Target* 30(1), 53-86.

Tummers, J., K. Heylen & D.Geeraerts. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus linguistics and linguistic theory* 1(2), 225-261.

Vandevoorde, L. (2016). *On semantic differences: a multivariate corpus-based study of the semantic field of inchoativity in translated and non-translated Dutch*. Doctoral dissertation, Ghent University.

The CRITT Translation Process Research Database. https://sites.google.com/site/centretranslationinnovation/tpr-db.

Translation Research in Corpora, Keystroke Logging and Eye Tracking (TRICKLET). http://www.anglistik.rwth-aachen.de/cms/Anglistik/Forschung/Laufende-Projekte/~gceg/Tricklet/?lidx=1.

Corpus-based contrastive studies: Beginnings, developments and directions

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

There is nothing new about the idea that languages can be compared, and multilingual texts have been with us since before the Rosetta Stone was inscribed. However, the field of corpus-based contrastive studies is still fairly young. In my talk I will outline the beginnings of corpus-based contrastive linguistics, take stock of recent developments and finally attempt to identify some prospects for the future of the field.

When multilingual corpora entered the scene in the early 1990s, they represented a new development in corpus linguistics, which had been predominantly monolingual until then, as well as in contrastive studies, which got a better empirical basis, new methods and a general boost. A seminal contribution was the compilation and completion of the English-Norwegian Parallel Corpus (ENPC) and its sister project, the English-Swedish Parallel Corpus (Aijmer et al. 1996). It has been pointed out that the advent of multilingual corpora greatly revived the latent interest in contrastive studies (Salkie et al. 1998) and “contributed to its awakening from an almost dormant state” (König 2012: 4).

At present, corpus-based contrastive linguistics is a well-established field of research which can be distinguished from the neighbouring fields of translation studies, learner corpus studies and typological studies (König 2012). In its essence, contrastive analysis is “the systematic comparison of two or more languages, with the aim of describing their similarities and differences” (Johansson 2007: 1). While contrastive analysis used to be associated with applied linguistics (ibid.; Salkie et al. 1998), the focus on immediate applications has now been toned down (Johansson 2012: 46). With the use of multilingual corpora, the systematic comparison will also facilitate “new insights into the languages compared – insights that are likely to be unnoticed in studies of monolingual corpora” (Aijmer & Altenberg 1996).

The plans for the English-Norwegian Parallel corpus were first presented as a project idea at the ICAME conference in 1993 (Johansson & Hofland 1994), and soon after the project was joined by other teams developing similar corpora of English/Swedish and English/Finnish (Aijmer et al. 1996). However, the revival of contrastive studies was not limited to the Scandinavian context. Other teams elsewhere started compiling parallel corpora of other language pairs, for example English and French in the PLECI corpus[1] . Clearly, the time was right for such an enterprise. The general technological advances made it possible to extend corpus methods to multilingual corpora, with the development of software for the alignment of translated texts and for parallel concordancing (Ebeling 2016). At the same time, corpus linguistics was becoming mainstream thanks to the spread of personal computers and the advent of the internet, which greatly facilitated access to corpora.

The corpus model represented by the ENPC can be termed a *bidirectional translation corpus*. This means that there are comparable original texts in both (all) languages concerned and translations into the other language(s). Importantly, the use of translations implies that the linguist need not determine beforehand which (pair/set of) linguistic items should be included in the comparison: instead, cross-linguistic correspondences can be established by exploiting *translation paradigms*, i.e. “the set of forms in the target text which are found to correspond to particular words or constructions in the source text” (Johansson 2007: 23). An example is the English speaking verb *talk*, which is translated predominantly by the Norwegian verb *snakke* but also with *prate* (‘chat’). *Snakke* in turn is regularly translated by either *talk* or *speak* (and both verbs have some less recurring members of their respective paradigms). Such use of the translations helps us identify relevant items for the cross-linguistic comparison. The bidirectional model is particularly well suited for contrastive studies because the translation relation can provide a *tertium comparationis* (a “background of sameness”) for the cross-linguistic

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/pleci.html.

comparison, at the same time as the presence of original texts in both languages provides a way of avoiding the translation bias which can be a problem for unidirectional translation corpora (Johansson 2007; Ebeling & Ebeling 2013). However, a challenge for the bidirectional model is the limited availability of translated texts for relevant language pairs and text types. Comparable corpora, i.e. corpora of original texts matched by criteria such as text type, publication date, and target audience, can provide more varied data, but do not have the in-built tertium comparationis of translation corpora. In my talk, I will give examples of studies based on both translation and comparable corpora.

In recent years the field of corpus-based contrastive linguistics has diversified to include comparisons of more languages and language pairs. More attention has also been given to genre and text types (e.g. Lefer & Vogeleer 2014) and to a greater variety of linguistic features at both micro-and macro-levels of analysis (as evidenced e.g. by papers published in *Languages in Contrast*). Nevertheless, some important challenges remain, for instance the availability of suitable data and reliable methods for making valid cross-linguistic comparisons based on parallel (translation) data as well as comparable data. In one of his last papers, Johansson (2012: 64f.) acknowledges the challenge that "we need to learn more about how we can best exploit multilingual corpora". Still, he argues that multilingual corpora, "if used with care and imagination, (…) lead us beyond what we know or did not see so clearly. This is the essence of the cross-linguistic perspective."

References

Aijmer, K., Altenberg, B. & Johansson, M. (eds). (1996). *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies*. Lund: Lund University Press.

Ebeling, J. (2016). Contrastive linguistics in a new key. *Languages in Contrast 20 Years on*. Special issue of *Nordic Journal of English Studies* 15(3), 7-14.

Ebeling, J. & Ebeling, S. O. (2013). *Patterns in Contrast.* Amsterdam & Philadelphia: Benjamins.

Johansson, S. (2007). *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies.* Amsterdam & Philadelphia: Benjamins.

Johansson, S. (2012). Cross-linguistic perspectives. In M. Kytö (ed.) *English Corpus Linguistics: Crossing Paths*. Amsterdam: Rodopi, 45-68.

Johansson, S & Hofland, K. (1994). Towards an English-Norwegian Parallel Corpus. In U. Fries, G. Tottie & P. Schneider (eds). *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora*, Zürich 1993. Amsterdam: Rodopi, 25-37.

König, E. (2012). Contrastive linguistics and language comparison. *Languages in Contrast* 12(1), 3-26.

Lefer, M.-A. & Vogeleer, S. (eds). (2014). *Genre- and Register-related Discourse Features in Contrast*. Special issue of *Languages in Contrast*, 14(1).

Salkie, R., Aijmer, K. & Barlow, M. (1998). Editorial. *Languages in Contrast* 1(1), v-xii.

Using corpora for evaluating translations and language change

Juliane House
Hamburg University, Hellenic American University
jhouse@fastmail.fm

My talk is in two parts. In the introductory first part I discuss the use and function of corpora in translation studies. In part two I present an example of a diachronic corpus-based translation project.

Corpora are today fruitfully used to 'lend an element of empirical inter-subjectivity to the concept of equivalence, especially if the corpus represents a variety of translators' (Altenberg & Granger 2002: 17). Translation corpora provide a useful methodological tool for clarifying hypothesized equivalences and for establishing reliable patterns of translation regularities (Krein-Kuehle 2013; Zanettin 2014).

Regardless of frequency and representativeness, corpus data are useful because they are quite often simply better data than those derived from accidental introspections. As in other lines of scientific inquiry, in corpus research it is important to assess the relative value of the analytical-nomological paradigm on the one hand, where already existing hypotheses (and categories) are to be confirmed or rejected, and where variables are explicated and operationalized, and the explorative-interpretative paradigm on the other hand, where in-depth case studies are conducted to develop categories for newly emerging phenomena. It is important that these two lines of inquiry, the quantitative and the qualitative, be not considered mutually exclusive, rather they should be regarded as supplementing each other.

In the second part of my talk I discuss the micro-diachronic, corpus-based project 'Covert Translation – Verdecktes Uebersetzen' which I directed at the Hamburg Research Centre on Multilingualism from 1999- 2012 (cf. Becher et al. 2009; Kranich et al. 2012). This project links qualitative work based on my translation evaluation model with quantitative analyses as well as ensuing re-contextualized case studies of the translation relation.

The general assumption underlying this project is that the dominance of the English language in many domains today leads to variation and change of indigenous communicative norms in German (and other languages) in both covert translations (House 2015) from English into German and original German texts, such that a gradual adaptation to Anglophone norms results. More concretely, we hypothesized that adaptations to Anglophone communicative norms can be located along dimensions of empirically established communicative preferences such as the ones established for German and English by myself (House 2006). An influence of English on German texts would manifest itself in the use of certain linguistic items and structures in German translations and comparable German texts in genres where Anglophone influence is particularly noticeable, such as popular science and economic texts.

To test the project hypothesis, we put together a corpus holding about 650 texts, featuring English original texts, their translations into German and comparable German originals as well as a few French and Spanish control texts. (These were later abandoned for research pragmatic reasons.) The genre 'popular science' comprises articles on topics of general sociopolitical relevance in two time frames: 1978-1982 and 1999-2002. These texts totaling about 700,000 words were selected from publications by official organs (e.g. *Scientific American* and *New Scientist* and their satellite journals in other languages). The genre 'economic texts' comprises around 300,000 words of annual reports by globally operating companies in the time frames 1978-1982 and 1999-2002 updated from 2002-2006 including letters to shareholders, missions, visions, corporate statements and product presentations. The fact that different time frames are included in the corpus enables evaluation of language variation and change along a time axis.

Since the three subcorpora differ substantially in terms of word count, we limited ourselves to presenting percentages and normalized frequencies.

For the qualitative analysis, House's translation evaluation model (latest version 2015) was used with some 80 English and German original and translated texts. The analyses showed that in the English texts, various linguistic means such as mental processes, simulated dialogues, evaluations, structural parallelism, framing and other narrative devices are used to personalize texts and establish a relationship with readers. The German texts avoid using these linguistic means in the first time frame, which makes them generally less interactional and person-oriented, with the result that readers are more instructed than entertained. We captured these differences in what we chose to call 'subjectivity' and 'addressee orientation'. Some of these differences were found to be less frequent in the second time frame.

The quantitative analyses were needed to verify the results of the qualitative analyses with regard to the diachronic development of the frequency of occurrence of expressions of subjectivity and addressee orientation found to be vulnerable to variation and change under Anglophone influence. They included personal pronouns, mental processes, expressions of epistemic modality and connectives.

In the third re-contextualization phase we looked at the translation relation in detail focusing on the above linguistic items across the two time frames in context, set at five preceding sentences and five ensuing ones. The linguistic items were extracted from the three subcorpora followed by manual annotation with a view to establishing co-occurrence and collocation patterns as well as syntactic and textual positions vis-à-vis the organization of information.

In the talk, I present examples of the behavior of personal pronouns, sentence initial connectives, epistemic modal marking and linking constructions including observed changes over time. Results confirm the continued operation of a "cultural filter" for certain phenomena thus blocking Anglophone impact while refuting it for others. Discussion and interpretation of the project results involve language-typological and socio-cultural factors as well as perceived formal and functional differences in the use of salient expressions.

References

Altenberg, B. & Granger, S. (2002). Recent Trends in Cross-Linguistic Lexical Studies. In B. Altenberg & S. Granger (eds). *Lexis in Contrast. Corpus-based Approaches*. Amsterdam: Benjamins, 3-48.

Becher, V., House, J. & Kranich, S. (2009). Convergence and Divergence of Communicative Norms through Language Contact in Translation. In K. Braunmueller & J. House (eds). *Convergence and Divergence in Language Contact Situations.* Amsterdam: Benjamins, 125-152.

House, J. (2006). Communicative Styles in English and German. *European Journal of English Studies* 10, 125-152.

House, J. (2015). *Translation Quality Assessment: Past and Present*. London: Routledge.

Kranich, S., J. House & V. Becher (2012). Changing Conventions in English and German Translations of popular Science Texts. In K. Braunmueller & C. Gabriel (eds). *Multilingual Individuals and Multilingual Societies.* Amsterdam: Benjamins, 315-335.

Krein-Kuehle, M. (2013). Towards a High-Quality Translation Corpus: The Cologne Specialized Translation Corpus. In M. Krein-Kuehle & U. Wienen (eds). *Koelner Konferenz zum Fachuebersetzen 2010.* Frankfurt/Main: Lang, 3-17.

Zanettin, F. (2014) Corpora in Translation. In J. House (eds). *Translation: A Multilingual Approach.* Basingstoke: Palgrave Macmillan, 178-199.

Expanding the third code: Corpus-based studies of constrained communication and language mediation

Haidee Kruger
Macquarie University, North-West University
haidee.kruger@mq.edu.au

Corpus-linguistic as well as computational research has yielded substantial evidence that translated texts demonstrate linguistic patterns that systematically distinguish them from non-translated texts in the same language (see Baroni & Bernardini 2006; Volansky et al. 2015; Zanettin 2013). These findings have provided support for the notion that translated language is a kind of "third code" (Frawley 2000 [1984]) shaped by the sociocognitive constraints that operate in mediating between two linguistic codes. The third code of translation is thus studied as a variety in its own right, with its own "standards and structural presuppositions and entailments" that arise from "the bilateral accommodation of a matrix and target code" (Frawley 2000 [1984]: 257). The linguistic patterns that most consistently appear to set translated language apart from non-translated language are typically framed as three tendencies:

(1) increased explicitness of lexicogrammatical encoding;
(2) a preference for comparably more conventional, conservative, or standard usage;
(3) cross-linguistic influence, priming or transfer, often of a subtle and indirect type leading to quantitative differences in linguistic patterning between non-translated and translated language

(see, for example, Delaere et al. 2012; Hansen-Schirra 2011; Hansen-Schirra et al. 2007; Lefer & Vogeleer 2013; Teich 2003; Xiao & Dai 2014).

There are some key unresolved issues in this area of research. First, the forces conditioning these three tendencies intersect and overlap. For example, the higher frequency of the use of optional complementiser *that* in English translations compared to English original writing has been ascribed to subconscious processes of explicitation as a result of the cognitive complexity of translation (Olohan & Baker 2000; Kruger & Van Rooy 2016b), in line with arguments that more analytical variants are preferred in more cognitively demanding processing contexts (Hawkins 2003; Mondorf 2014; Rohdenburg 1996). It has also been ascribed to cross-linguistic influence from source languages with obligatory complementisers (Becher 2010), and to a risk-avoidant, conservative overadjustment to the norms for formal writing (Kruger in press b; Kruger & De Sutter in press). A significant challenge is therefore to disentangle the various explanatory hypotheses proposed for the features of translated language (or demonstrate their interaction), in order to provide more robust and parsimonious explanations for the observed distinctive patterning of translations (see House 2008; Malmkjær 2005).

A second challenge has been accounting for the fact that a diverse set of factors other than translated versus non-translated status plays a role in conditioning the frequency and use of a linguistic feature, including user variables like writer or translator background, discursive variables like register (and sub-register), and text-internal lexicogrammatical variables (see De Sutter et al. 2017; Neumann 2011). These factors introduce intersecting and co-varying dimensions of variability that have, until recently, not been adequately accounted for in comparisons of translated and non-translated language, despite the existence of substantive traditions of research that focus exactly on these dimensions of variability, for example in register studies and variationist linguistics.

Lastly, and most pertinently for this paper, it has long been suggested that the features of translated language are not unique to translations only, but are evident in a larger set of varieties characterised by diverse communicative constraints. These constraints include, amongst others, discourse production under conditions of bi- or multilingual language activation, and the relaying or "mediation" of an existing message. In this view, the features that typify translations are reframed more broadly as features or "universals" of language mediation, language contact, bi- or multilingual discourse production, or constrained communication (see, for example, Bisiada 2017; Chesterman 2004, 2014; Gaspari & Bernardini 2010; Granger 2018; Kajzer-Wietrzny 2018; Kruger 2012; Kruger & Van Rooy 2016a; Lanstyák & Heltai 2012; Shlesinger & Ordan 2012; Steiner 2008, Ulrych & Murphy 2008).

How to investigate and account for linguistic variability across various dimensions simultaneously is the key question that informs current corpus-based studies of constrained communication and language mediation. These studies are highly interdisciplinary, drawing together research on translation, interpreting, editing, second-language varieties, and learner varieties, with the aim of seeking a more general explanation for the linguistic similarities among these varieties, while remaining attuned to their differences. In this process, there has been an increasing move towards multifactorial corpus analysis methods, to address the full range of factors that potentially condition variation in multiple varieties that are subject to similar and distinct constraints.

In this paper, I survey existing research on this topic, pointing out the key theoretical assumptions and focusing on the methodological approaches used thus far. I set out an argument for modelling the recurrent features of different forms of constrained or mediated communication conceptually and empirically as sets of overlapping "varioversals", extending the concept of varioversals put forward by Szmrecsanyi & Kortmann (2009 and elsewhere) to refer to linguistic features that typify varieties of language that share certain constraints. I argue that constrained varieties may be seen as probabilistically conditioned by five overarching and interacting constraint dimensions (conceived as continua rather than binaries), enabling us to model the similarities and differences between varieties:

(1) Language activation (monolingual—bilingual)
(2) Modality and register (spoken—written—multimodal)
(3) Text production (independent/unmediated—dependent/mediated)
(4) Proficiency (native/proficient—non-native/learner)
(5) Task expertise (expert—non-expert)

I focus on the variationist, multifactorial and interdisciplinary corpus-linguistic approach required to disentangle these five constraint dimensions, and discuss some recent studies to illustrate how these methods, combined with the theoretical assumptions of a coherent explanatory framework of constrained communication, may assist in unscrambling some of the motivations behind the features that typify constrained varieties, highlighting their similarities and differences. My focus will be specifically on two sets of empirical work. I will outline corpus-based work on editing and revision, considering methods for investigating how editing reshapes published texts. I reflect on how a research agenda on editing connects with the notion of constrained language, by demonstrating how translation and editing are similar and different as a consequence of differences in the constraints that operate on them, and how editing affects translations, as well as the language production of L1 and L2 writers in differential ways (see Bisiada 2017, 2018a, 2018b; Kruger 2012, 2017, in press a; Kruger & Van Rooy 2017).

Subsequent to this, I discuss studies that investigate translation alongside other contact-influenced varieties, including second-language varieties of English, contact-influenced first-language varieties, learner language, and bilingual language production. I focus on recent work utilising state-of-the-art multifactorial methods to investigate (a) the *that*/zero alternation in written contact and non-contact varieties (Kruger in press b; Kruger & De Sutter in press) and (b) aggregate patterns of register variation across contact varieties (Kruger & Van Rooy 2018). I conclude by highlighting limitations and desiderata for work in this area.

References

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

Becher, V. (2010). Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1), 1-28.

Bisiada, M. (2017). Universals of editing and translation. In S. Hansen-Schirra, O. Czulo & S. Hofmann (eds). *Empirical Modelling of Translation and Interpreting*. Berlin: Language Science Press, 241-275.

Bisiada, M. (2018a) Editing nominalisations in English−German translation: When do editors intervene? *The Translator* 24(1), 35-49.

Bisiada, M. (2018b). Translation and editing: A study of editorial treatment of nominalisations in draft translations. *Perspectives: Studies in Translation Theory and Practice* 26(1), 24-38.

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

Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, K. Malmkjær & D. Gile (eds). *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*. Amsterdam & Philadelphia: John Benjamins, 1-13.

Chesterman, A. (2014). Translation Studies Forum: Universalism in Translation Studies. *Translation Studies* 7, 82-90.

De Sutter, G., Lefer, M.-A. & Delaere, I. (2016). Introduction. In G. De Sutter, M.-A. Lefer & I. Delaere (eds). *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin: De Gruyter Mouton, 1-8.

Delaere, I., De Sutter, G. & Plevoets, K. (2012). Is translated language more standardised than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target* 24(2), 203-224.

Frawley, W. (2000 [1984]). Prolegomenon to a theory of translation. In L. Venuti (ed.) *The Translation Studies Reader*. London & New York: Routledge, 250-263.

Gaspari, F. & Bernardini, S. (2010). Comparing non-native and translated language: Monolingual comparable corpora with a twist. In R. Xiao, ed. *Using Corpora in Contrastive and Translation Studies.* Newcastle: Cambridge Scholars Publishing, 215-234.

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.

Hansen-Schirra, S. (2011). Between normalization and shining-through: Specific properties of English-German translations and their influence on the target language. In S. Kranich, V. Becher, S. Höder & J. House (eds). *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*. Amsterdam & Philadelphia: John Benjamins, 133-162.

Hansen-Schirra, S., Neumann, S. & Steiner, E. (2007). Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in Contrast* 7(2), 241-265.

Hawkins, J. (2003). Why are zero-marked phrases close to their heads? In G. Rohdenburg & B. Mondorf, B. (eds). *Determinants of Grammatical Variation in English*. Berlin: Mouton de Gruyter, 175-204.

House, J. (2008). Beyond intervention: Universals in translation? *trans-kom* 1, 6-19.

Kajzer-Wietrzny, M. (2018) Interpretese vs. non-native language use: The case of optional *that*. In M. Russo, C. Bendazzoli & B. Defrancq (eds). *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, 97-113.

Kruger, H. (2012). A corpus-based study of the mediation effect in translated and edited language. *Target* 24(2), 355-388.

Kruger, H. (2017). A corpus-based study of the effects of editorial intervention: Implications for the features of translated language. In G. De Sutter, M.-A. Lefer & I. Delaere (eds). *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin: De Gruyter Mouton, 113-156.

Kruger, H. (In press a). Does editing matter? Editorial work, endonormativity and convergence in written Englishes in South Africa. In R. Hickey (ed.) *English in Multilingual South Africa*. Cambridge: Cambridge University Press.

Kruger, H. (In press b). *That* again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*.

Kruger, H. & De Sutter, G. (In press). Alternations in contact and non-contact varieties: Reconceptualising *that*-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition and Behavior*.

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.

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.

Kruger, H. & Van Rooy, B. (2017). Editorial practice and the progressive in Black South African English. *World Englishes* 36(1), 20-41.

Kruger, H. & Van Rooy, B. (2018). Register variation in written contact varieties of English: A multidimensional analysis. *English World-Wide* 39(2), 214-242.

Lanstyák, I. & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures* 13(1), 99-121.

Lefer, M.-A. & Vogeleer, S. (eds). (2013). *Interference and Normalisation in Genre-controlled Multilingual Corpora.* Amsterdam & Philadelphia: John Benjamins.

Malmkjær, K. (2005). Norms and nature in translation studies. *Synaps* 16, 13-19.

Mondorf, B. (2014). (Apparently) competing motivations in morpho-syntactic variation. In B. MacWhinney, A. Malchukov & E. Moravcsik (eds). *Competing Motivations in Grammar and Usage*. Oxford: Oxford University Press, 209-228.

Neumann, S. (2011). *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. Berlin: Mouton de Gruyter.

Olohan, M. & Baker, M. (2000). Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1(2), 141-158.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2), 149-182.

Shlesinger, M. & Ordan, N. (2012). More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target* 24(1), 43-60.

Steiner, E. (2008). Empirical studies of translations as a mode of language contact: "Explicitness" of lexicogrammatical encoding as a relevant dimension. In P. Siemund & N. Kintana (eds). *Language Contact and Contact Languages*. Amsterdam & Philadelphia: John Benjamins, 317-346.

Szmrecsanyi, B. & Kortmann, B. (2009). Vernacular universals and angloversals in a typological perspective. In M. Filppula, J. Klemola & H. Paulasto (eds). *Vernacular Universals and Language Contact: Evidence from Varieties of English and Beyond.* New York: Routledge, 33-53.

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

Ulrych, M. & Murphy, A. (2008). Descriptive translation studies and the use of corpora: Investigating mediation universals. In C. Taylor Torsello, K. Ackerley & E. Castello (eds). *Corpora for University Language Teachers.* Bern: Peter Lang, 141-166.

Volansky, V., Ordan, N. & Wintner, S. (2015). On the features of translationese. *Literary and Linguistic Computing*, 30(1), 98-118.

Xiao, R. & 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.

Papers and posters

Translation directionality and expertise: An empirical approach

Laura Aguiar de Souza Penha Marion, Gaëtanelle Gilquin, Marie-Aude Lefer
Université catholique de Louvain
laura.aguiar@uclouvain.be, gaetanelle.gilquin@uclouvain.be, marie-aude.lefer@uclouvain.be

This presentation describes the main methodological, descriptive and theoretical objectives of a new research project in the field of Empirical Translation Studies.

The project aims at investigating two key constructs in Translation Studies, namely translation directionality (L1 vs. L2 translation) and expertise (novice vs. expert translation), through the prism of a range of purported translation properties. We will work on the French-English language pair, looking at two aspects that are essential to language usage, viz. production (as represented in computerized corpora) and processing (as visible through e.g. keyboard logging).

The field of Empirical Translation Studies (ETS) (Ji 2016; De Sutter et al. 2017) is made up of two complementary, but usually separate, research strands, which rely on different types of data: process- and product-oriented ETS. Translation process research (O'Brien 2011) traditionally relies on think-aloud protocols, keyboard logging and eye tracking. It mainly focuses on cognitive aspects of translation (e.g. development of translation competence). In process-based studies, little attention is paid to the linguistic properties of the product (i.e. the translated text itself). Product-oriented research, by contrast, typically relies on corpora. This approach, known as Corpus-Based Translation Studies (CBTS; Olohan 2004), uses corpus data drawn from monolingual comparable corpora (original and translated texts in Language$_X$) or parallel corpora (source texts in Language$_X$ and corresponding target texts in Language$_Y$) with a view to identifying what distinguishes translated language from non-translated, original language. In-depth linguistic analyses of translations are the norm in CBTS, but the field is currently facing issues of data interpretation because corpora offer no access to the translation process (e.g. problem-solving strategies). In this project we aim to show that ETS can greatly benefit from a genuine combination of the product- and process-oriented approaches. This type of rapprochement has already been advocated and successfully implemented in monolingual usage-based linguistics, which recommends methodological pluralism as a way of approaching the different facets of language usage (cf. Ellis & Simpson-Vlach 2009; Gilquin & Gries 2009; Schönefeld 2011).

The first construct that will be investigated is directionality, which refers to the direction of translation, i.e. whether translation is done into the translator's first language (L2>L1) or into his/her second/foreign language (L1>L2) (cf. Campbell 1998; Grosman et al. 2000; Adab 2005; Pokorn 2005, 2011). Two main assumptions are found in the literature on directionality: (i) translation competence is asymmetrical (cf. Beeby Lonsdale 2009), L2 translation being of lower quality than L1 translation (e.g. Durban 2011), and (ii) L2 translation is cognitively more challenging/demanding than L1 translation (Chang 2011). To date, however, directionality has been under-researched in ETS (Apfelthaler 2013; Hunziker Heeb 2016), especially in CBTS, for lack of suitable corpus data (only L1 translation is represented in the corpora traditionally used in CBTS).

The second construct explored in this project is translation expertise. The comparison of novice translation (translation by translator trainees) and expert/professional translation is one of the leading topics in process-oriented studies, where the acquisition of translation competence features high on the research agenda (e.g. Englund Dimitrova 2005). However, these studies being based on small amounts of data, the generalizability of their findings tends to be rather weak. In CBTS, by contrast, the degree of generalizability tends to be higher but the focus has mainly been on professional translation (with a few exceptions, such as Redelinghuys & Kruger 2015 and Castagnoli 2016). Expertise, like directionality, still awaits solid empirical, usage-based investigation.

As will be shown in the presentation, we aim to study directionality and expertise empirically, using product and process data representing the French-English language combination (in both translation directions). The product data will be taken from the *Multilingual Student Translation* corpus (MUST; Granger & Lefer 2017) and from a corpus of professional translations, corresponding to novice and expert product data respectively. The process data will be collected through screen recording and keyboard logging. These data will be analysed through the prism of the typical features of translated language, i.e. so-called translation universals (Baker 1993, 1995; Laviosa 2002; Olohan 2004), such as explicitation/increased explicitness, lexico-syntactic simplification and normalization. The universal status of these translation features is quite controversial nowadays (cf. Becher 2010) and has made way for the concept of translation properties (Neumann 2011). In addition, several scholars have argued that, rather than being typical of translation alone, these properties are common to different forms of mediated language and constrained communication, such as learner language and other instances of bilingual communication (Lanstyák & Heltai 2012). Taking these new insights into account, we would also like to approach translation properties from a different angle. So far, these properties have been operationalized by means of linguistic indicators that can be extracted from corpora fairly easily and (semi-) automatically (e.g. explicitation: connectors and optional *that*; simplification: type-token ratio, lexical density and average sentence length) (see e.g. Redelinghuys & Kruger 2015; Bernardini et al. 2016). In this project, we will refine these operationalizations (e.g. by exploring the use of new indicators) and go beyond indicators that can be automatically extracted, notably thanks to the computer-aided manual annotation of some translation properties. To do so, we will rely on a new translation-oriented annotation system (TAS) developed within the framework of the MUST project (Granger & Lefer 2017).

Our presentation will zoom in on the main objectives of the project:
(1) Theory and description: qualify claims on translation directionality and expertise by describing, on the basis of solid empirical evidence, the typical properties of L1/L2 and novice/expert translation, comparing output produced by translator trainees and professional translators working out of and into their mother tongue (here French><English). By going beyond the impressionistic descriptions found in the literature, we will be able to refine our understanding of translation as a form of mediated language and constrained communication and thus build bridges between Translation Studies and fields like Second Language Acquisition, Learner Corpus Research and Contact Linguistics, which are all concerned with different types of mediated language.
(2) Methodology: test the potential of product-process/linguistic-cognitive triangulation in ETS, when starting from the linguistic product (i.e. the translated text) and using cognitive process data as a supplement to shed light on the features of the product; develop new operationalizations for translation properties, combining linguistic indicators that can be (semi-)automatically extracted from corpora and manually annotated translation features.

References

Adab, B. (2005). Translating into a Second Language: Can We, Should We? In G. M. Anderman & M. Rogers (eds). *In and Out of English: For Better, For Worse?* Clevedon: Multilingual matters, 227-241.

Apfelthaler, M. (2013). *Directionality research in translation and interpreting studies (1/2): A shorter than short history*, http://cogtrans.net/blog.htm.

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: Benjamins, 233-250.

Baker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research. *Target* 7(2), 223-243.

Becher, V. (2010). Abandoning the Notion of "Translation-Inherent" Explicitation. Against a Dogma of Translation Studies. *Across Languages and Cultures* 11(1), 1-28.

Beeby Lonsdale, A. (2009). Directionality. In M. Baker & G. Saldanha (eds). *Routledge Encyclopedia of Translation Studies*. London & New York: Routledge, 84-88.

Bernardini, S., Ferraresi, A. & Miličević, M. (2016). From EPIC to EPTIC. Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28(1), 58-83.

Campbell, S. (1998). *Translation into the Second Language*. London & New York: Longman.

Carl, M. (2012). Translog-II: A Program for Recording User Activity Data for Empirical Reading and Writing Research. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA).

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, 1-21.

Chang, V. C.-Y. (2011). Translation directionality and the Revised Hierarchical Model: An eye-tracking study. In S. O'Brien (ed.) *Cognitive Explorations of Translation*. UK: Continuum, 154-174.

De Sutter, G., Lefer, M.-A. & Delaere, I. (2017). *Empirical Translation Studies: New methodological and theoretical traditions*. Trends in Linguistics. Studies and Monographs. Berlin: De Gruyter.

Durban, C. (2011). *Translation – getting it right. A guide to buying translation*. American Translators Association, http://www.atanet.org/publications/Getting_it_right.pdf.

Ellis, N.C. & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5(1), 61-78.

Englund Dimitrova, B. (2005). *Expertise and Explicitation in the Translation Process*. Amsterdam & Philadelphia: John Benjamins.

Gilquin, G. & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1), 1-26.

Granger, S. & M.-A. Lefer (2017). Bridging the gap between learner corpus research and translation studies: The Multilingual Student Translation corpus. *4th Learner Corpus Research Conference*, Bolzano, 5-7 October 2017.

Grosman, M., Kadric, M., Kovačič, I. & Snell-Hornby, M. (eds). (2000). *Translation into non-mother tongues: In professional practice and training*. Tubingen: Stauffenburg.

Hunziker Heeb, A. (2016). Professional translators' self-concepts and directionality: indications from translation process research. *The Journal of Specialized Translation* 25, 74-88.

Ji, M. (ed.) (2016). *Empirical Translation Studies: Interdisciplinary Methodologies Explored*. Sheffield: Equinox.

Lanstyak, I. & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures* 13(1), 99-121.

Laviosa, S. (2002). *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam & New York: Rodopi.

Neumann, S. (2011). *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: De Gruyter.

O'Brien, S. (ed.) (2011). *Cognitive Explorations of Translation*. UK: Continuum.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London: Routledge.

Pokorn, N. K. (2005). *Challenging the Traditional Axioms. Translation into a non-mother tongue*. Amsterdam & Philadelphia: John Benjamins.

Pokorn, N. K. (2011). Directionality. In Y. Gambier & L. van Doorslaer (eds). *Handbook of Translation Studies: Volume 2*. Amsterdam & Philadelphia: John Benjamins, 37-39.

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.

Schönefeld, D. (ed.) (2011). *Converging Evidence: Methodological and Theoretical Issues for Linguistic Research*. Amsterdam: John Benjamins.

Questions in translations from English into Swedish and Norwegian

Karin Axelsson
University of Gothenburg, University of Skövde
karin.axelsson@his.se

The aim of this study is to establish what happens to questions (here defined as clauses/fragments followed by a question mark) in English fiction texts when they are translated into Swedish and Norwegian (*bokmål*). The focus is on non-congruent translations (cf. Fredriksson 2016), in particular when there is a change into another question type or into a non-question, e.g. a statement. Swedish and Norwegian are closely related languages with similar word order: they belong to the SVO group but have the finite in second position (V2). Being North Germanic languages, Swedish and Norwegian are also fairly closely related to English: the major question types — *wh*-questions, *yes/no*-questions, alternative questions, declarative questions, tag questions, fragments and indirect questions — are common in all three languages. The default translation is thus a congruent translation, exemplified by the polar questions in (1) and the *wh*-questions in (2).

(1) *Is everything ready?*
ST: *Är allt klart?*
NT: *Er alt klart?* (AH)

(2) *When was that?*
ST: *När var det?*
NT: *Når var det?* (JSM)

It is therefore interesting to investigate when and why non-congruent translations are used: are these due to subtle contrastive differences, e.g. one language preferring a particular question type more than the other languages, or to translation universals such as explicitation and normalisation, or to the translator's choice of an idiomatic clause/phrase instead of a less idiomatic congruent translation?

There have been many translation/contrastive studies investigating English vs. Swedish (e.g. Aijmer 2017) and Norwegian (e.g. Ebeling & Ebeling 2017), but questions in general have seemingly not been treated except for a restricted pilot study on Norwegian vs. English by Wikberg (1996), who, however, disregarded question fragments by only looking at sentences of at least 7-10 words before the question mark. There are also previous studies only dealing with tag questions in translations between English and Swedish/Norwegian (Axelsson 2006, 2009).

Here are some research questions within this project:

- To what extent are phrases/clauses ending in questions marks not translated into phrases/clauses ending in question marks?
- To what extent and why are various question types translated non-congruently? What question types (or other clause types) are these non-congruent translations?
- How often and why are there changes in polarity in the translations of questions?
- What differences in translation patterns of questions can be seen between Swedish and Norwegian?

The data for this study has been retrieved from the *English-Swedish Parallel Corpus* (Aijmer et al. 2001) and the *English-Norwegian Parallel Corpus* (Johansson et al. 1999/2002). These corpora, the ESPC and the ENPC, are similarly designed and may be searched using the same web interface. For English originals (EO), the ESPC comprises 25 English novel extracts with their translations into Swedish and the ENPC 30 English novel extracts with their translations into Norwegian. The 24 files in EO that are (in principle) identical to both corpora are used in the present study. These subcorpora are here called ESPC-EO24 and ENPC-EO24, each containing about 330,000 words; they are jointly referred to as ENPC/ESPC-EO24. In both ESPC-EO24 and ENPC-EO24, all question marks were retrieved with contexts and translations. After minor differences due to scanning errors have been dealt with, the dataset consists of 2,028 English questions. The dataset of Swedish translations of English questions is called ESPC-ST24 and the dataset of Norwegian translations of English questions ENPC-NT24. However, all question marks have also been retrieved from the full subcorpora ESPC-ST24 and ENPC-ST24, i.e. including the use of question marks in Swedish/Norwegian when there is no question mark in English. This reveals that the total number of question marks is just slightly lower in ESPC-ST24 than in ESPC/ENPC-EO24 (2,012) and very close to ESPC/ENPC-EO24 in ENPC-NT24 (2,025). This might lead to the conclusion that there are no large differences between the originals and the translations as to question marks, but there are changes in two ways: question marks disappear in translations in some cases and are added in other cases.

A question mark in English is not translated into a question mark in 4-5% of the cases; here, full stops are predominant, but other punctuation marks such as commas and exclamation marks also occur. There are also a few cases where the translator, for some reason or other, has not translated the question at all, and therefore some potential question marks have disappeared in the translations. English *wh*-questions (n=663) are not translated into Swedish and Norwegian *wh*-questions in 7-8% of the cases: in such cases, most translations are instead *yes/no*-questions or declaratives, but there are also e.g. *wh*-fragments and indirect questions. English *yes/no*-questions (n=604) are not translated into Swedish and Norwegian *yes/no*-questions in 9% of the cases: all the major question types are found as translations here but declaratives, *wh*-questions and indirect questions prevail. Various kinds of fragments form about a quarter of all questions in ENPC/ESPC-EO24, most of them being *wh*-fragments and NP-fragments. As to *wh*-fragments (n=179), there is a significant difference between translations into Swedish and Norwegian (p<0.01): 31% are not translated into Norwegian *wh*-fragments and 47% are not translated into Swedish *wh*-fragments. This is partly due to English fragments with *what about* and *what if* having counterparts in Norwegian but not in Swedish. In (3), Swedish uses a complete *wh*-question instead:

(3) *What about that computer?* (DF)
NT: *Hva med den datamaskinen?*
ST: *Hur är det med den där datorn?*
How is it with that computer

A clear majority of English declarative questions are not translated into declarative questions in Swedish and Norwegian, as in (4), where the Scandinavian languages use polar questions.

(4) *You know the family?* (SG)
ST: *Känner du familjen?*
NT: *Kjenner du familien?*
Know you the family

A change of polarity is not uncommon, especially in the translations of English negative questions where 11-12% become positive in Norwegian and Swedish, as in (5). By comparison, only 2% of positive questions become negative in translation.

(5) *How could I not have been happy for him?* (JB)
ST: *Hur kunde jag vara annat?*
NT: *Hvordan kunne jeg være annet?*
How could I be otherwise

References

Aijmer, K. (2017). The semantic field of obligation in an English-Swedish contrastive perspective. In K. Aijmer & D. Lewis (eds). *Contrastive analysis of discourse-pragmatic aspects of linguistic genres*. Cham: Springer, 13-32.

Aijmer, K., Altenberg, B. & Svensson, M. (2001). English-Swedish Parallel Corpus: Manual. Available at https://sprak.gu.se/english/research/research-activities/corpus-linguistics/corpora-at-the-dll/espc, accessed 22 January 2018.

Axelsson, K. (2006). Tag questions in English translations from Swedish and Norwegian – are there differences? In B. Englund Dimitrova & H. Landqvist (eds). *Svenska som källspråk och målspråk: aspekter på översättningsvetenskap [Swedish as a source language and as a target language: aspects on translation studies].* Göteborg: Göteborgs universitet, 4-21.

Axelsson, K. (2009). Tag questions in translations between English and Swedish. In B. J. Epstein (ed.) *Northern lights: translation in the Nordic countries*. Oxford: Peter Lang, 81-106.

Ebeling, S. O. & Ebeling, J. (2017). A cross-linguistic comparison of recurrent word combinations in a comparable corpus of English and Norwegian fiction. In M. Janebová, E. Lapshinova-Koltunski & M. Martinková (eds). *Contrasting English and other languages through corpora*. Cambridge: Cambridge Scholars Publishing, 2-31.

Fredriksson, A.-L. (2016). *A corpus-based contrastive study of the passive and related constructions in English and Swedish*. PhD thesis. Gothenburg: Department of Languages and Literatures, University of Gothenburg.

Johansson, S., Ebeling J. & Oksefjell, S. (1999/2002). English-Norwegian Parallel Corpus: Manual. Available at http://www.hf.uio.no/ilos/english/services/omc/enpc, accessed 22 January 2018.

Wikberg, K. (1996). Questions in English and Norwegian: evidence from the English-Norwegian Parallel Corpus. In C. E. Percy, C. F. Meyer & I. Lancashire (eds). *Synchronic corpus linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16).* Amsterdam: Rodopi, 17-28.

Primary sources

AH Hailey, A. (1984). *Strong medicine.* London: Michael Joseph.
ST: Hailey, A. (1984). *Stark medicin.* Transl. by S. J. Lundvall. Stockholm. Bonnier.
NT: Hailey, A. (1985). *Sterk medisin*. Transl. by A. S. Seeberg. Oslo: Dreyers.

DF Francis, D. (1989). *Straight.* London: Michael Joseph.
ST: Francis, D. (1991). *Dödligt arv.* Transl. by H. Pettersson. Stockholm: Wahlström & Widstrand.
NT: Francis, D. (1991). *Dødelig arv.* Transl. by H. Kolstad. Oslo: Gyldendal.

JB Barnes, J. (1991). *Talking it over.* London: Jonathan Cape.
ST: Barnes, J. (1992). *Tala ut*. Transl. by Ingvar Skogsberg. Stockholm: Forum.
NT: Barnes, J. (1993). *En trekanthistorie.* Transl. by K. Ofstad. Oslo: H. Aschehoug & Co.

JSM Smiley, J. (1991). *A thousand acres.* London: Flamingo HarperCollins.
ST: Smiley, J. (1998). *Tusen tunnland.* Transl. by Y. Stålmarck. Stockholm: Norstedts.
NT: Smiley, J. (1992). *Fire tusen mål.* Transl. by A. Elligers. Oslo: J. W. Cappelens.

SG Grafton, S. (1990). *"D" is for deadbeat.* London: Pan Books.
ST: Grafton, S. (1993). *D som i drunknad*. Transl. by B. Crona. Höganäs: Mysterious Press.
NT: Grafton, S. (1993). *"D" for druknet.* Transl. by I. Rogde. Oslo: Tiden Norsk Forlag.

Characterizing absence: A corpus-based contrastive study of verbless sentences in English and Russian

Antonina Bondarenko
Université Paris Diderot – Paris 7
tonyabondarenko@gmail.com

The present paper presents a corpus-based quantitative analysis of semantic and pragmatic factors associated with the absence of a verb from a sentence. Profound cross-linguistic differences make it particularly relevant to compare English with Russian, a language known for permitting the most liberal use of verbless sentences among the Indo-European family and possessing a highly developed morphological case system and flexibility of word order. In contrast, English is known for its dependence on the finite verb phrase, the lack of a zero-copula construction and register restrictions on verbal ellipsis (McShane 2000; Kopotev 2007; Stassen 2013). The aim of the paper is to statistically characterize verbless sentences in Russian and English in terms of key semantic and pragmatic elements, contextual features, and verbal translation patterns. What statistically key factors distinguish verbless sentences from verbal sentences and how do they compare in English and Russian? What does a quantitative analysis of the translations of verbless sentences reveal about the verbs that are pragmatically implicated in verbless sentences? Finally, what does cross-linguistic comparison of verbless sentences suggest about the notion of predication?

The analysis is based on three types of corpora. First, a comparable corpus, which includes several works of dialogue-based realist fiction and plays in English and Russian,[1] is used to statistically analyze each text independently and the exposed features of verbless sentences are compared cross-linguistically. Secondly, multiple translations of these works[2] are aligned with the originals to form a parallel text sub-corpus which is used to explore translation patterns bi-directionally. We use the parallel corpus to examine the verbal correspondences of non-antecedent based verbless sentences and, following the contrastive linguistics principles of Guillemin-Flescher (2003) and parallel corpus criteria of Stolz (2007) and Nádvorníková (2017), we look for verbal translation patterns which recur across the different translators, texts, and translation directions. We also use it to establish the rate at which verbs are gained, or lost, in translation. Thirdly, following Baker (1993, 2000) and Zanettin (2012), we add a third-language sub-corpus consisting of Russian and English translations from French[3] and attempt to control for source language influence in the translation data.

Most previous studies of verbless sentences have used corpora for illustrative purposes, frequently basing the analysis on fragmented data and often privileging a syntactic discussion (McShane 2000; Weiss 2013; Zidane 2014). The difficulty of detecting absence automatically has prevented verbless sentence studies from taking a corpus-linguistic approach (Weiss 2011: 139). A survey of existing parsed corpora reveals that verb-centric syntactic modeling and fixed annotation tagsets very often hamper automatic verbless sentence extraction (Landolfi et al. 2010). The absence of a zero-marker is identified as one of the current limits of corpus annotation (Loock 2016: 33). Overcoming these problems, we have developed an alternative method of automatic verbless sentence extraction with a recall of 95%.

Using specific sentence segmentation and the Trameur annotation, alignment and statistical text analysis software (Fleury & Zimina 2014), verbless sentences were automatically extracted, submitted to various statistical analyses and aligned with several translations. Trameur permits automatic correction of a large portion

[1] Dostoevskij, *Brat'ja Karamazovy* (1880); Orwell, *Animal Farm* (1945); Bradbury, *Fahrenheit 451* (1953); Solženicin, *Olen' i Šalašovka* (1953); Pinter, *The Caretaker* (1960); Strugackij & Strugackij, *Piknik na Obočine* (1972); Franzen, *The Corrections* (2001); Prilepin, *San'kja* (2006).

[2] Šinkar', *451° po Farengejtu* (1956); Bethell & Burg, *The Love-Girl and the Innocent* (1969); Bouis, *Roadside Picnic* (1977); Polock, *Skotnyj Dvor* (1980); Task, *Skotskij Ugolok* (1989); Pevear & Volokhonsky, *The Brothers Karamazov* (1990); Avsay, *The Karamazov Brothers* (1994); Pribylovskij, *Zverskaja Ferma: Skazka* (2002); Summ, *Popravki* (2005); Doroševič, *Storož* (2006); Bormashenko, *Roadside Picnic* (2012); Gusev, Parker & Ryabovolova, *Sankya* (2014).

[3] Gilbert, *The Stranger* (1946); Adamovich, *Neznakomec* (1966); Gal, *Postoronnij* (1968); Ward, *The Stranger* (1988).

of morphosyntactic tagging errors and also makes it possible to visualize verbless sentences aligned with multiple translations in their original context. Statistical analysis included the computation of characteristic elements, repeated segments and collocations in terms of key forms, lemmas and grammatical categories associated with verbless sentences, as compared with a verbal reference corpus. Verbless utterances and their translation correspondences were then manually annotated for antecedent-based verbal ellipsis, discourse type, the presence of a verb and its tense. The results also include the rates at which translation correspondences are verbalized.

Preliminary results reveal that markers of deixis and informal speech statistically characterize verbless sentences in both languages. These include deictic particles *это (èto, this/it)* and *вот (vot, here)* in Russian and *this* in English, interjections *oh* in English and *ну (nu, well)* in Russian, and the Russian second-person singular pronoun *ты (ty*, the familiar form of *you)*. The latter suggests that verbless sentences are associated with informally addressing the interlocutor. These results expose a statistical pragmatic restriction on the use of verbless sentences, since the interpretation of these markers requires an established common ground between the speaker and addressee. The requirement of a common ground is also exposed in the strong correlation of verbless sentences with direct speech as opposed to narration. Such results are in line with qualitative semantic analyses that link certain types of verbless sentences to the utterance situation (e.g. Selivërstova 1973; Paillard 1984).

Verbless sentences seem to be statistically associated with emphasis in Russian. Repeated segments calculation reveals that the pattern in which the deictic *это (èto, this/it)* is followed by emphatic particles, such as *ведь (ved', after-all)* and *уж (už, indeed),* is overrepresented in Russian verbless sentences. Example (1) illustrates the pattern and shows a verb phrase used to create the emphasis in English.

(1) *А ведь непредвиденное-то обстоятельство – это ведь я!*
a ved' nepredvidennoe-to obstojatel'stvo – èto ved' ja!
PART PART unforeseen.that-ADJ.PART circumstance-NOM – this-DEM after.all-PART I-NOM
'My word, unforeseen circumstances – he means me!'

The frequency difference between Russian and English verbless sentences is statistically significant and not explainable by syntactic factors. Non-antecedent based verbless utterances dominate in both languages. Despite Russian formally allowing more productive verbal ellipses, virtually all of the antecedent-based Russian ellipses were matched by ellipses in English translation correspondences and a higher frequency of antecedent-based ellipses was surprisingly found in English. Translation correspondences suggest that half of the Russian verbless utterances become verbal in English translation. Finally, verbless sentence frequency differences are revealed to be greater than the difference in the general frequency of verbs in the two languages.

Our quantitative analysis exposes the importance of considering verbless sentences from a pragmatic perspective. We emphasize that verb-centric definitions of predication are inadequate not only for practical corpus annotation purposes (Landolfi et al. 2010), but also theoretically due to cross-linguistic instability in contrastive analysis. We argue for the traditional notion of a predicate as a syntactic link (Creissels 1995) and for the need to separate it from the semantic notion of predication.

References

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

Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator. *Target* 12(2), 241-266.

Creissels, D. (1995). *Éléments de Syntaxe Générale.* Paris: Presses Universitaires de France.

Fleury, S. & Zimina, M. (2014). Trameur: A framework for annotated text corpora exploration. In L. Tounsi & R. Rak (eds). *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin: Dublin City University and Association for Computational Linguistics, 57-61.

Guillemin-Flescher, J. (2003). Théoriser la traduction. *Revue Française de Linguistique Appliquée* 8(2), 7-18.

Kopotev, M. (2007). Where Russian syntactic zeros start: Approaching Finnish? *Slavica Helsingiensia* 32, 116-137.

Landolfi, A., Carmela S. & Voghera, M. (2010). Verbless clauses in Italian, Spanish and English: A Treebank annotation. In S. Bolasco, I. Chiari & L. Giuliano (eds). *Proceedings of JADT 2010, the 10th International Conference on Statistical Analysis of Textual Data*. Rome: CISU, 1187-1194.

Loock, R. (2016). *La Traductologie de Corpus.* Villeneuve d'Ascq: Presses Universitaires du Septentrion.

McShane, M. (2000). Verbal ellipsis in Russian, Polish and Czech. *The Slavic and East European Journal* 44(2), 195-233.

Nádvorníková, O. (2017). Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela [Online]* HS-21, 1-28.

Paillard, D. (1984). *Énonciation et Détermination en Russe Contemporain*. Paris: Institut d'études slaves.

Selivërstova, O. (1973). Semantičeskij analiz predikativnyx pritjažatel'nyx konstrukcij s glagolom byt'. *Voprosy Jazikoznanija* 5, 95-105.

Stassen, L. (2013). Zero copula for predicate nominals. In M. S. Dryer & M. Haspelmath (eds). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Stolz, T. (2007). *Harry Potter* meets *Le Petit Prince*: On the usefulness of parallel corpora in crosslinguistic investigations. *STUF-Sprachtypologie und Universalienforschung* 60(2), 100-117.

Weiss, D. (2011). Bezglagol'nye konstrukcii russkoj razgovornoj reci: Ix tipologija i status v lingvisticeskom opisanii. In I. M. Boguslavskij, L. L. Iomdin & L. P. Krysin (eds)., *Slovo i jazyk: Sbornik statej k vos'midesjatiletiju Ju. D. Apresjana*. Moskva: Jazyki slavjanskix kul'tur, 139-155.

Weiss, D. (2013). The lazy speaker and the fascination of emptiness: Colloquial Russian from a typological perspective. In I. Kor-Chahine (ed.) *Current Studies in Slavic Linguistics*. Amsterdam: John Benjamins, 91-123.

Zanettin, F. (2012). *Translation-driven Corpora*. Manchester: St Jerome Publishing.

Zidane, R. (2014). The Use of Verbless Sentences in English Literature. *Ulakbilge* 2(4), 57-76.

A rather interesting topic: A contrastive study of English *rather*, Dutch *eerder* and French *plutôt*

Lieselotte Brems[1], Lobke Ghesquière[2], Gudrun Vanderbauwhede[2]
ULiège[1], UMONS[2]
lbrems@uliege.be, lobke.ghesquiere@umons.ac.be, gudrun.vanderbauwhede@umons.ac.be

In Germanic and Romance languages, originally temporal constructions have developed attested preferential or contrastive uses and degree uses, e.g. English *rather*, Dutch *eerder*, German *eher* and French *plutôt*, Italian *piuttosto* respectively. These adverbial forms have developed from comparative forms of temporal or speed adjectives (OE *hraeþ-er*, D *eer-(d)er*, G *e-(h)er*, FR *plus-tôt*, IT *piu-tosto*). Despite the remarkable cross-linguistic morphosyntactic and pragmatic-semantic similarity of these forms, to our knowledge no thorough comparative study of these expressions has been carried out so far. This paper aims to go some way in filling this gap and reports on the synchronic study of three of these adverbial markers – English *rather*, Dutch *eerder* and French *plutôt*. The a, b, and c examples in (1) to (3) illustrate their temporal, preference and degree uses respectively. Today, Dutch *eerder* is the only adverb to still have temporal uses (2a). *Rather* lost its temporal use in the Middle English period (Rissanen 2008). In French, the univerbated form *plutôt* is generally restricted to express contrastive and degree meanings. The temporal meaning is restricted to the comparative two-word construction *plus tôt* 'more early'. The *Trésor de la Langue Française informatisé* (TLFi) does mention instances in which *plutôt* is used with a temporal meaning, as in (3a), but lists them as 'ungrammatical' or 'wrong'. It will be interesting to see if such uses are attested in our (translated) data.

(1) English
a. *I sawe the Heauen and the Starres..neither rather or later to rise or go downe.* (OED, s.v. rather I.2.a)
b. *Unfortunately, this adequacy was a reminder that his problem has not been his lack of style but rather his abundance of insincerity* (Rissanen 2002: 357)
c. *Sachs understood that she was playing with him but he rather enjoyed the way she went about it.* (OED, s.v. rather, 6b)

(2) Dutch
a. *Waarom werd dat niet eerder aan de orde gesteld?* (DBNL)
'Why wasn't this matter brought up sooner?'
b. *Maar ik noem dat geen fatalisme, ik zie er eerder een vorm van verweer in.* (DBNL)
'But I wouldn't call it fatalism. I'd rather see a kind of defence in it'
c. *En hoewel de Zweden van nature uit een eerder stijf en nauwgezet volk zijn kan men er dan ook in elke krantenkiosk, en gewoon tentoongesteld, tijdschriften zien met foto's die onze zedenmeesters de kolieken zouden doen krijgen.* (DBNL)
'And even though the Swedes are by nature a rather stiff and meticulous people one can find displayed in every newspaper stand magazines with pictures that would give our moralists the gripes'

(3) French
a. *Arriver plutôt ou plus tard.* (TLFi, s.v. plutôt, adv. A.1)
'To arrive sooner or later.'
b. *Il ne peut plus supporter cette présence, ce mystère derrière la porte; plutôt la mort tout de suite, si c'est elle, que l'angoisse de l'inconnu.* (TLFi, s.v. plutôt B.1)
'He can no longer bear this presence, this mystery behind the door; rather death right now, if that is what it is, than the anxiety of the unknown.'
c. *Une personne plutôt jolie.* (TLFi, s.v. plutôt C.1)
'A rather pretty person.'

Both *rather* and *plutôt* have already been the object of study in the grammaticalization literature. Traugott & König (1991: 203-204), for instance, argue that both developed along a pathway leading from temporal to preferential meaning. For English, Rissanen (2008) argued that in the Modern English period the preferential or contrastive reading made way for the degree modifying reading, constituting a pathway from the Old English temporal meaning to contrastive meaning to degree modifying reading. For French, Mokni (2008) posits a shift from temporal to contrastive only, making no mention of the degree modifying use. A quick corpus search, however, returns many degree uses and these uses are also mentioned in established dictionaries. Recently, Ghesquière & Brems (2017) hypothesized on the basis of 20th century data, a similar grammaticalization path for Dutch *eerder*, leading from temporal to contrastive to degree uses, yet with the observation that English seems to have progressed further along the cline, as Dutch *eerder*, unlike *rather*, still has temporal uses and is used less often as a degree modifier than its English counterpart.

We will extend the synchronic contrastive English-Dutch study of Ghesquière & Brems (2017) to include French *plutôt* and draw up typologies of the different uses of these adverbial markers and compare them qualitatively and quantitatively, both in terms of their semantics/pragmatics and their structural behaviour. Semantico-pragmatically we will try to come to a fine-grained classification of the different preferential and degree uses. For the preferential uses, our starting-point is Quirk et al.'s (1985: 638-639) classification of textual relations, including (pure) contrast, reformulation and replacement. For the degree uses, attention will go to the upscaling or downscaling nature of intensification and to the specific structural type of intensification scale involved (Kennedy & McNally 2005). Structurally, the specific complementation patterns in which the adverbs engage is one of the parameters that will be looked into, both language-specifically and across translations.

This study is based on both original and translated language, which allows us to assess the degree of intertranslatability of the three constructions, similarities or differences in usage and frequency in original and translated text as well as to come to a better understanding of the different language-internal uses of *rather*, *eerder* and *plutôt*. The monolingual English corpus used is the British books section of the *WordbanksOnline* corpus. For Dutch, we queried the twentieth century texts of the *Digitale Bibliotheek voor de Nederlandse Letteren* [Digital Library of Dutch Literature] (DBNL). For French, data were extracted from *Frantext*. The translation data for this study are extracted from the bi-directional *Dutch Parallel Corpus* (DPC), a 10-million-word parallel corpus comprising texts in Dutch, English and French with Dutch as a pivotal language.

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 at http://www.frantext.fr/.
WB: *WordbanksOnline*. Available online at https://wordbanks.harpercollins.co.uk/.

References

Ghesquière, L. & Brems, L. (2017). Time, preference and intensity: A contrastive study of *rather (than)* and *eerder (dan)*. Paper presented at ICAME38: Corpus et Orbis: Interpreting the World through Corpora, 24-28 May 2017, Charles University, Prague.

Juge, M.L. (2002). Unidirectionality in grammaticalization and lexical shift: The case of English *rather*. *Berkeley Linguistics Society 28*, 147-154.

Kennedy, C. & L. McNally. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2), 345-381.

Mokni, M. (2008). La grammaticalisation de l'adverbe *plutôt* et l'évolution du système grammatical. *Linx* 59, 171-184.

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

Rissanen, M. (1999). On the adverbialization of RATHER: Surfing for historical data. In H. Hasselgard & S. Oksefjell (eds). *Out of corpora: Studies in honour of Stig Johansson* (Language and Computers 26). Amsterdam: Rodopi, 49-59.

Rissanen, M. (2008). From 'quickly' to 'fairly': On the history of *rather*. *English Language and Linguistics* 12(2), 345-359.

Traugott, E.C. & E. König. (1991). The semantics-pragmatics of grammaticalization revisited. In E.C. Traugott & B. Heine (eds). *Approaches to Grammaticalization*, Amsterdam: Benjamins, 189-218.

Comparative constructions in French-speaking Belgian learners of English: A contrastive approach

Zoé Broisson[1], Kristel Van Goethem[1,2]
Université catholique de Louvain[1], F.R.S.-FNRS[2]
zoe.broisson@student.uclouvain.be, kristel.vangoethem@uclouvain.be

This study provides a contrastive analysis of comparative constructions in French and English, on the one hand, and investigates the acquisition of these constructions by French-speaking Belgian learners of English, on the other. A difference is made between pupils following the Content and Language Integrated Learning (CLIL) method and pupils enrolled in the traditional language learning settings.

CLIL is one of the leading didactic methods to have been developed to promote multilingualism through education in Europe. This approach involves the teaching of content school subjects through the medium of a target language distinct from the school's mainstream language (Eurydice 2012). Although the CLIL method has been extensively documented internationally (Ruiz de Zarobe 2008; Rumlich 2016; Seikkula-Leino 2007), its impact on second or foreign language acquisition remains a subject of scholarly debate.

Assessing the impact of CLIL is the purpose of an on-going large-scale longitudinal and interdisciplinary research project in French-speaking Belgium (cf. Hiligsmann et al. 2017). The present study contributes to this line of research by contrasting the use of comparative constructions in native French and native English and by investigating CLIL and non-CLIL pupils' use of English comparative constructions.

We first investigate the similarities and differences between French and English through a contrastive analysis with comparative constructions as a *tertium comparationis*. We base this analysis on the description of French and English 'ordinary' comparative constructions, illustrated in (1-3), in grammars by Biber et al. (1999), Grevisse (1975), Riegel et al. (1994) and Quirk et al. (1978), leaving aside 'idiomatic' comparatives such as the construction *may/might as well* (Sawada 2007). Ordinary comparatives have been extensively studied within the frameworks of syntax and semantics (Bresnan 1973; Fuchs et al. 2008), and cognitive or functional typology (Andersen 1983; Heine 1997). The contrastive analysis reveals that French and English have similar syntactic (and irregular) ways of marking comparison at their disposal, such as the construction [Modifying adverb + Adjective/Adverb/Phrase (+correlative construction)] (examples (1-2)). However, English presents an additional morphological alternative to mark the comparative of superiority, in the form of the suffix *–er* (3).

(1) *Snow White is more beautiful (than the queen)*
(2) *Blanche Neige est plus belle (que la reine)*
(3) *Snow White is taller (than the queen)*

We hypothesize that this discrepancy between the two languages acts as a potential obstacle to the correct use of comparative constructions by our population of French-speaking learners of English. From a usage-based perspective, such as the one adopted in Ellis and Cadierno (2009), the main challenge faced by learners consists in the competition between the specific constructions of their native and foreign language. In light of our contrastive analysis, we formulate the following two research questions:

1. Which formal types of comparative constructions do learners use in their L1 and L2 productions?
2. Do CLIL learners develop a more native-like use of comparative constructions (in terms of formal types of constructions, diversity and accuracy) thanks to more target language input?

We propose to answer these research questions through the extraction and the analysis of 399 instances of comparative constructions in five small-scale comparable datasets: one control corpus of L1 English, two corpora of L1 French (CLIL and non-CLIL) and two corresponding learner corpora of L2 English (CLIL and non-

CLIL). The data constituting the French L1 and English L2 corpora were collected in 2017 among 438 sixth-grade secondary school pupils averaging 18.5 years old and attending the schools involved in the Belgian project. The data constituting the English L1 corpus were collected in 2016 among 70 university students from Florida (USA), averaging 19.4 years old. The five corpora used in this study were compiled according to the same criteria, and contain data that result from the production of a same writing task (email to a friend) averaging 303.76 words in length. Basing ourselves on results obtained in a study of global complexity measures conducted on the writing of the same population of pupils in 2016 (Bulon et al. 2017), we estimate the English proficiency level of our pupils to be ranging from B1 to B2 on the CEFR scale (Council of Europe 2001).

When examining the distribution of syntactic, morphological and irregular comparatives in the datasets, our preliminary results reveal, as expected, that both CLIL and non-CLIL learners show a tendency to use less morphological comparatives in English than the native speakers from the control corpus. Therefore, the results of our analysis suggest that comparative constructions represent an area of difficulty for French-speaking learners of English from Belgium. This is in part because learners have to overcome their preference for syntactic constructions likely due to the influence of their mother tongue, but also because of other factors transpiring from our corpus analysis, such as the overall higher productivity of comparative constructions in English as compared to French in the student populations under study.

Through the analysis of the errors made by our learners, we also identify that the two groups of pupils diverge with regards to the frequency and the type of error related to the use of comparative constructions. We distinguish three types of errors: (1) Functional errors, due to confusion between comparative and superlative marking; (2) Formal errors consisting in the use of a syntactic comparative in domains taking morphological marking; and (3) Syntactic errors, due to the addition or omission of syntactic elements within the comparative construction. We observed that the non-CLIL learners made errors in all three categories, whereas the CLIL learners only made errors of the functional type.

The results of the corpus analysis also highlight that CLIL pupils generally form English comparative constructions more frequently and more diversely than non-CLIL pupils, but non-CLIL pupils use comparatives in proportions and in collocations that are more native-like. Both the CLIL and the traditional teaching methods each have their advantages (and drawbacks), but ultimately we argue that the CLIL approach produces better learning outcomes in the case of comparative constructions, because of a lower rate of error and a higher degree of diversity.

References

Andersen, P. (1983). *Word Order Typology and Comparative Constructions.* Amsterdam: John Benjamins.

Biber, D., Johansson, S. & Leech, G. (1999). *Longman Grammar of Spoken and Written English*. 4th impr. Harlow: Pearson education.

Bresnan, J. (1973). Syntax of the Comparative Clause Construction in English. *Linguistic Inquiry* 4(3), 275-343.

Bulon, A., Hendrikx, I., Meunier, F. & Van Goethem, K. (2017). Using global complexity measures to assess second language proficiency: Comparing CLIL and non-CLIL learners of English and Dutch in French-speaking Belgium. *Papers of the Linguistic Society of Belgium* 11(1), 1-25.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Ellis, N. & Cadierno, T. (2009). Constructing a Second Language. *Introduction to the Special Section. Annual Review of Cognitive Linguistics* 7, 111-139.

Eurydice. (2012). *Content and Language Integrated Learning (CLIL) at school in Europe. Belgium – French community: national description – 2004/5.* Retrieved from https://publications.europa.eu/en/publication-detail/-/publication/1aa4cf19-aefa-47bf-9b55-21bb3a65f9eb/language-en/format-PDF/source-62563919 [Last accessed 3-01-2018].

Fuchs, C., Fournier, F. & Le Goffic, P. (2008). Structures à subordonnée comparative en français: Problèmes de représentations syntaxiques et sémantiques. *Lingvisticæ Investigationes* 31(1), 11-61.

Grevisse, M. (1975). *Le bon usage: Grammaire française avec des remarques sur la langue française d'aujourd'hui*. Gembloux: Duculot.

Heine, B. (1997). *Cognitive Foundations of Grammar.* Oxford: Oxford University Press.

Hiligsmann, P., Van Mensel, L., Galand, B., Mettewie, L., Meunier, F., Szmalec, A., Van Goethem, K., Bulon, A., De Smet, A., Hendrikx, I. & Simonis, M. (2017). Content and Language Integrated Learning: Linguistic, Cognitive and Educational Perspectives. *Cahiers du Girsef* 109, 1-25.

Quirk, R., Greenbaum, S. & Leech, G. (1978). *A grammar of contemporary English.* London: Longman.

Riegel, M., Pelat, J. & Rioul, R. (1994). *Grammaire Méthodique du Français*. Paris: Quadrige.

Ruiz de Zarobe, Y. (2008). CLIL and foreign language learning: a longitudinal study in the Basque country. *International CLIL Research Journal* 1, 60-73.

Rumlich, D. (2016). *Evaluating bilingual education in Germany: CLIL students' general English proficiency, EFL self-concept and interest.* Frankfurt am Main: Lang.

Sawada, O. (2007). The cognitive patterns of construals in comparatives. In M. Nakano & K. Park (eds). *Proceedings of the 10th Conference of Pan-Pacific Association of Applied Linguistics*, Tokyo: PAAL, 209-226.

Seikkuna-Leino, J. (2007). CLIL learning: achievement levels and affective factors. *Language and Education* 21(4), 328-341.

Translating punctuation: A corpus-based proposal to enhance students' output

Paola Brusasco, Elisa Corino
Università di Torino
paola_brusasco@yahoo.it, elisa.corino@unito.it

Punctuation has been receiving renewed attention, presumably because of its somewhat unstable norms and, even more so, because of the shift in style deriving from its minimalist and/or altered use in contemporary media and social media discourse.

While overall aware of the punctuation system of their own language, students seem to consider it mainly for its prosodic or "elocutional function" (Crystal 2015: 69). Therefore, in the translation class attention needs to be drawn to three main aspects: 1) the ways in which punctuation contributes to meaning-making – and is therefore an integral part of the translation process; 2) the conventions determining its usage on the basis of text type and medium in each of the languages involved in the translation process, and 3) how to 'translate' punctuation.

Originally, punctuation carried out the primary function of providing indications for reading (intonation, pauses and prosody). Today, it is clear that punctuation works at a level which goes beyond prosodic and strictly syntactic phenomena. Punctuation should therefore be analyzed as one of the constituents of textuality, considering its role in organizing information and structuring it (Ferrari & Lala 2013; Ferrari 2017; Bertuccelli Papi 2017).

Text, communication and cognition are the keywords to describe forms and functions of punctuation, as punctuation delimits units of locutionary, illocutionary and perlocutionary value, contributes to building the text structure and identifying the focus of the sentence, and guides the reader in decoding the hierarchy of information.

Still, students seem to rely mostly on prosodic criteria or produce 'punctuation calques' which do not correspond to either syntactic rules or communicative purposes. Examples from students' translations include

Fortunately, a combination of flexible exchange rates, strong international reserves, better monetary regimes, and a shift away from foreign-currency debt provides some measure of protection.
Fortunatamente, la combinazione tra tassi di cambio flessibili, forti riserve internazionali, migliori regimi monetari e un allontanamento dal debito in valuta estera fornisce una certa tutela.

and

Indeed, some emerging markets – for example, Colombia – had been issuing public debt at record-low interest-rate spreads over US treasuries.
Sicuramente, alcuni mercati emergenti, per esempio la Colombia, hanno continuato ad emettere debito pubblico a differenziali di tassi d'interesse bassissimi sui titoli del Tesoro statunitense.

Here the English use of the comma is simply transferred to the Italian text with an awkward outcome, as the sentence takes on an ambiguous – if not different – illocutionary value.

While such punctuation calques can also work in the opposite direction, with students' texts written in English following the Italian practice of usually not separating the linker from the main body of the sentence through a comma, as in "Therefore it is necessary to integrate technology into the learning process", for the purpose of the present study only translation activities have been considered.

From a contrastive point of view there are differences in the use and distribution of punctuation in Italian as against other languages (Ferrari & Stojmenova 2015; Buzzoni 2008). With reference to texts translated from English into Italian, the most frequent cases are the controversial use of the single dash, the serial comma, the comma splice, or the use of a comma after a linking word at the beginning of a sentence or clause (Corino 2015, 2017).

The present contribution reports on a teaching experience carried out in two MA-level courses (English Linguistics and Translation Theories and Practices) at the Department of Foreign Languages and Literatures and Modern Cultures of the University of Torino. Our aim is to ascertain if and to what extent a small, specially created, parallel corpus of about 31,000 words illustrating the use of specific punctuation items can help students to use them effectively and correctly in their own translations. The selected items are the comma after a linking word and the single dash.

To do so, students were divided into two groups: the first translated two LSP texts (astronomy) from English into Italian with the help of a reference table summarizing the contrastive uses of punctuation in the two languages. This is our control group.

The other group started by creating a parallel corpus of articles in English with their Italian translations, taken from *Scientific American* and *Le Scienze* respectively. Then they did some corpus-based exercises with the aim to identify the patterns of use – if any – of the items we selected. Finally, they translated the same texts that were previously given to the control group.

The translations have been collected and are being reviewed to test our hypothesis that the inclusion of punctuation-oriented corpora in the education of future translators (or language specialists equipped with translation skills) will enhance the students' awareness of the role of punctuation in the meaning of texts and enable them to use it effectively and correctly according to text type and target language conventions.

References

Baker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research. *Target* 7(2), 223-243.

Bertuccelli Papi, M. (2017). Naturalezza e marcatezza nella punteggiatura inglese. In A. Ferrari, L. Lala, F. Pecorari (a cura di), *L'interpunzione oggi (e ieri). L'italiano e le altre lingue europee*. Firenze: Cesati, 265-284.

Buzzoni, M. (2008). La punteggiatura nei testi di lingua inglese. In B. Mortara Garavelli (a cura di), *Storia della punteggiatura*. Roma-Bari: Laterza, 442-491.

Corino, E. (2015). Connettivi pragmatici e virgole. Descrizione di un pattern in incipit di enunciato, *RiCognizioni* 2/4 (2015), 11-25.

Corino, E. (2017). Connettivi e virgola: tradurre la punteggiatura tra attrito, norma e uso. Paper presented at SLI 2017, Napoli 27-29 Settembre 2017.

Crystal, D. (2015). *Making a Point. The Pernickety Story of English Punctuation*. London: Profile Books.

Fantinuoli, C., Zanettin, F. (eds). (2015). *New directions in corpus-based translation studies*. Berlin: Language Science Press.

Ferrari, A. (2017). La punteggiatura italiana oggi. Un'ipotesi comunicativo-testuale. In A. Ferrari, L. Lala, F. Pecorari (a cura di), *L'interpunzione oggi (e ieri). L'italiano e le altre lingue europee.* Firenze: Cesati, 19-36.

Ferrari, A. & Lala, L. (2013). La virgola nell'italiano contemporaneo. Per un approccio testuale (più) radicale, *Studi di grammatica italiana XXIX-XXX*, 479-501.

Ferrari, A. & Stojmenova, R. (2015). Virgole tedesche e virgole italiane a confronto, tra teoria e descrizione, *RiCognizioni* 2/4, 27-44.

Oakes, M. P., Meng, J. (eds). (2012). Quantitative Methods in Corpus-based Translation Studies. Amsterdam, PA: John Benjamins.

Cross-linguistic perspectives on intensification in speech: A comparison of L1 French and L2 English and Dutch

Natacha Buntinx[1,2], Kristel Van Goethem[1,3]
Université catholique de Louvain[1], University of Oslo[2], F.R.S.-FNRS[3]
natacha.buntinx@student.uclouvain.be, kristel.vangoethem@uclouvain.be

Following the framework of contrastive linguistics (especially Gast 2012), intensification of adjectives is taken as a *tertium comparationis* that allows a function-based approach of how different languages hypothetically use different linguistic forms to express the same ontological category. We study how intensification of adjectives is used and formally realised in the trilingual output of native French speakers in spoken French, non-native English and non-native Dutch. Learner language is considered as language showing features of the L1 of the learner (Gast 2012: 7), which allows for an assessment of the extent to which L2 users are able to use foreign languages with the same diversity of constructions and of semantic nuances as in their L1, while using constructions that are more typical of the L2.

Intensification has been extensively described in terms of its semantic (Paradis 2001 on boundedness) and formal properties in English (Bolinger 1972; Ito & Tagliamonte 2003), Dutch (Broekhuis 2017; van der Wouden & Foolen 2017), and French (e.g. Riegel et al. 1994). These studies show that languages use a wide variety of constructions to express intensification, but that language-specific preferences can be observed (cf. Rainer 2015). Also, intensification has been studied from a second language acquisition perspective for non-native English (Granger 1998; Lorenz 1999; De Haan & van der Haagen 2012) and for non-native Dutch (Hendrikx et al. 2017), with a focus on learner writing. However, to our knowledge, no study of the kind has been conducted on speech corpora and no study took a cross-linguistic view on the semantic types of intensification into account. Beyond this, the contribution aims at exploring the effect of Content and Language Integrated Learning (CLIL) on the use of intensification. More specifically, we address the following research questions:

i. Which formal types of intensifying constructions do learners use in their L1 and L2s?
ii. Which semantic types of intensification do learners use in their L1 and L2s?
iii. Is there a difference between CLIL and non-CLIL learners in that regard?

The study examines these questions on the basis of a speech corpus (conversations about a holiday trip or a party). The corpus consists of spoken data from 64 students in the 6th grade from three different secondary schools and was collected in 2017, within the framework of an ongoing research project on CLIL in French-speaking Belgium (Hiligsmann et al. 2017). It represents a total of 16 English conversations (12 pupils in non-CLIL and 19 in CLIL), and 17 Dutch conversations (20 non-CLIL and 13 CLIL), comparable to similar data from the same students in L1 French.

Based on the literature review, we have created a typology covering different intensifying constructions – syntactic (e.g. degree adverbs), morphological (e.g. prefixes), lexical (e.g. strong adjectives), and prosodic constructions (e.g. strong accent) – and semantic types: boosters, which denote a high intensity (e.g. *really*); maximizers, which denote the range of degrees situated at the upper extremity of the scale (e.g. *extremely*); totality markers, which express the reaching of a maximal endpoint (e.g. *totally*); and undetermined markers, which intensify the adjectives to some undetermined extent (e.g. *how* in *how tall is that guy!*).

The typology was then tested against the data, and compared to a similar analysis of L1 spoken English data from *SACODEYL European Youth Language* and L1 spoken Dutch data from the *Corpus Gesproken Nederlands* (CGN). The data was annotated in EXMARaLDA (Schmidt & Wörner 2012).

Interlanguages tend to show less balanced proportions in terms of construction types than in L1 French. L2 Dutch and L2 English show different patterns of intensification, with a significant role played by prosody in

L2 English, especially in non-CLIL contexts where this strategy might be compensating for a lack of vocabulary (1), whereas it is almost absent from L2 Dutch.

(1) no it's **impóssible**

With respect to the semantic analysis, it was found that students do not use the different types of semantic intensification similarly in their native language and in the interlanguages. In fact, both interlanguages tend to demonstrate less diversity in terms of semantic types of intensification, and more importantly, students intensify to a more neutral level via boosters in interlanguages, as can be seen in (2) and (3), than in L1 French, in which they prefer to maximize meaning (4).

(2) well but there are . **very** nice thing to do there hm like hm . biking
(3) wat is **heel** belangrijk voor jij? // Alcohol of euh [lit. 'what is **very** important for you? // Alcohol or uh']
(4) avec notre chambre d'hôtel on a une **hyper** belle vue / sur euh sur toute la plage [lit. 'with our hotel room we have a **hyper** beautiful view / on uh on the whole beach']

König (2017) points out that "ultimately, an evaluative utterance of this kind typically tells us more about the speaker than about 'reality'" (König 2017: 30) and that "evaluations with intensifiers are extremely context-dependent" (König 2017: 28). Our comparison of the 'intensifying strength' by the same speakers in their L1 and L2s sheds new light on the importance of subjectivity and context in the analysis of intensification. The comparison with L1 Dutch and English allowed for some hypotheses in that regard. As L1 Dutch presents similar proportions of intensification strength as L1 French, the results in L2 Dutch cannot be said to be influenced by the target language. They might be related to the formal type of input received at school: pupils might be less familiar with more informal and expressive intensifiers used in youth language. In L2 English, on the other hand, the observed proportions are similar to those found in L1 English, indicating a possible influence of the target language, which does however not exclude explanations related to the more generalized use of all-round boosters (such as *very*) in the interlanguage.

CLIL and non-CLIL settings do not present significant differences in terms of formal and semantic types of intensification. The only aspect that seems impacted is the acquisition of specific forms: CLIL learners show a more diverse range of adverbials.

References

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

Bolinger, D. (1972). *Degree Words*. The Hague: Mouton.

Broekhuis, H. (2017). Modification by an intensifier. Taalportaal (3.1.2). http://taalportaal.org/taalportaal/topic/link/syntax__Dutch__ap__a3__a3_Modification.3.1.2.xml (accessed 02 November 2017).

De Haan, P. & van der Haagen, M. (2012). Modification of adjectives in very advanced Dutch EFL writing: A development study. *The European Journal of Applied Linguistics and TEFL* 1(1), 129-142.

Gast, V. (2012). Contrastive linguistics: Theories and methods. In B. Kortmann & J. Kabatek (eds)., *Dictionaries of Linguistics and Communication Science: Linguistic Theory and Methodology*. Berlin: Mouton de Gruyter.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. In A. P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Clarendon Press, 145-160.

Hendrikx, I., Van Goethem, K., Meunier, F. & Hiligsmann, Ph. (2017). Language-specific tendencies towards morphological or syntactic constructions: A corpus study on adjective intensification in L1 Dutch, L1 French and L2 Dutch. *Nederlandse Taalkunde [Dutch Linguistics]* 22(3), 389-420.

Hiligsmann, Ph., Van Mensel, L., Galand, B., Mettewie, L., Meunier, F., Szmalec, A., Van Goethem, K., Bulon, A., De Smet, A., Hendrikx, I. & Simonis, M. (2017). Content and Language Integrated Learning: linguistic, cognitive and educational perspectives. *Cahiers du Girsef* 10, 91-25.

Ito, R. & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society* 32(2), 257-279.

König, E. (2017). The comparative basis of intensification. In M. Napoli & M. Ravetto (eds)., *Exploring Intensification. Synchronic, diachronic and cross-linguistic perspectives.* Amsterdam & Philadelphia: John Benjamins Publishing Company, 15-32.

Lorenz, G. R. (1999). *Adjective Intensification: Learners Versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam & Atlanta: Rodopi BV.

Paradis, C. (2001). Adjectives and boundedness. *Cognitive Linguistics* 12(1), 47-65.

Rainer, F. (2015). Intensification. In P. O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (eds). *Word-Formation: An International Handbook of the Languages of Europe*. Berlin/Boston: De Gruyter Mouton, 1340-1351.

Riegel, M., Pellat, J. & Rioul, R. (1994). *Grammaire Méthodique du Français*. Paris: Presses universitaires de France.

Schmidt, T. & Wörner, K. (2012). EXMARaLDA. In J. Durand, G. Ulrike & G. Kristoffersen (eds)., *Handbook on corpus phonology*. Oxford: Oxford University Press, 402-419.

Wouden, van der T. & Foolen, A. (2017). A most serious and extraordinary problem. Intensification of adjectives in Dutch, German, and English. *Leuvense Bijdragen* 101, 82-100.

Accuracy in consecutive interpreting: The effect of experience and source text difficulty

Hanne Cardoen
Université de Mons
Hanne.CARDOEN@umons.ac.be

This PhD project focuses on the possible influence between note-taking, source-text difficulty and the fluency and accuracy of the target text. In the literature, much has been written on different note-taking systems and practical guidelines have been proposed for interpreting students (e.g. Rozan 1956; Van Hoof 1962; Matyssek 1989). However, the number of empirical studies on note-taking and consecutive interpreting remains limited and, moreover, many quality parameters have not yet been clearly defined. This research project therefore tries to shed some light on the characteristics of efficient notes in consecutive interpreting, building on Helle Dams (2007) work on accuracy, but also takes into account a number of other relevant factors, such as source-text difficulty and experience. Furthermore, both fluency and accuracy will be analyzed as two important subparameters of quality.

A corpus of thirty consecutive interpretations from English into Dutch has been collected, together with the corresponding corpus of thirty note-sets and fifteen retrospective interviews with the interpreters. In other words, the concept of 'corpus' is in this study applied as a collection of experimental data, as is common in Interpreting Studies (e.g. Dam 1996; Tang & Dechao 2016). This approach allowed for a triangulation of the data. Fifteen interpreters with different levels of experience, five professional interpreters, five advanced students and five novices, all interpreted two source texts with varying levels of difficulty. In regard to the methodology used, the variable **source-text difficulty** has been analyzed by applying triangulation, as the two texts were prepared intuitively, then analyzed objectively by studying parameters such as word frequency, lexical diversity, lexical density or syntactic complexity and finally were checked for subjective perception of the interpreters in a pilot study. The **notes** have been analyzed from a product-based angle, according to four parameters, namely note quantity, the number of full words, of abbreviations and of symbols. A process-based approach using a Smartpen has not been applied, as the data were collected several years ago and the technology at that time, unfortunately, still presented some flaws. **Fluency** has been analyzed by means of a number of variables (speech rate, disfluencies, filled pauses, etc.), the biggest hurdle being silent pauses. A pilot study (Cardoen & D'Amelio 2013) on the perception of disruptive silent pauses allowed to fix two different thresholds for grammatical and non-grammatical pauses which were higher than most thresholds proposed in the literature, but which might also provide a higher reliability. Finally, the variable of **accuracy** has been assessed by two judges, both native speakers of Dutch, who compared the transcripts of the six source and target texts and used a holistic rating scale from 1 to 5. The raters were provided with clear guidelines in order to limit the possibility that they would interpret categories or tasks differently. Both raters therefore discussed the rating scale and the criteria they would apply in advance and agreed on the importance of certain error types, in other words, they applied the same criteria. This accuracy assessment has turned out to be a reliable tool, as it has been compared in a pilot study with a more detailed approach using a grid which indicates the different types of errors and their relative importance or weight. Both approaches produced the same results. Nevertheless, the rating scale method is less time-consuming and more transparent for future analyses.

This presentation will discuss the effect of experience and source-text difficulty level on interpreting accuracy. First of all, the effect of the rising difficulty level on the fifteen interpreters was rather clear and straightforward, as 12 of the 15 subjects got a lower accuracy score for their interpretation of the difficult speech. The only exceptions were Master two-students. Moreover, during the easier source text, the professional interpreters were more accurate than the advanced students who interpreted more accurately than the beginning students. Surprisingly, this was not the case for the more challenging speech, as on the contrary, the Master two-students were then the most accurate interpreters, followed by the professional interpreters and, finally, the Master one-students. One of the five professional interpreters even figured amongst the five most inaccurate subjects when interpreting the complicated speech. This result is not in line with research on expertise, as experts especially outperform novices on difficult tasks, rather than in routine situations (Ericsson 2000: 210).

Nevertheless, accuracy is of course only one of the many quality parameters meaning that the isolated parameter in this study, accuracy, should not be misinterpreted as a measure of overall quality. It is possible that the professionals obtain a higher overall quality, when other parameters are also taken into account, such as voice, target language quality or accent, which were, however, beyond the scope of this study. Nevertheless, the quality assessments presented here oblige us to highlight that, at the end of their training, advanced interpreting students obtain accuracy scores which are comparable to those of professional interpreters. Though overall quality research studying only professional interpreters is of course preferable for the sake of validity, these advanced students might be representative subjects for research that focuses solemnly on accuracy.

This is especially the case for studies on consecutive interpreting, as several professionals highlighted during the retrospective interviews that they only rarely used the consecutive mode and mainly worked as simultaneous interpreters, which might also partly explain why the students obtained higher accuracy scores during the more challenging task. On the other hand, as the results are not in line with research on expertise (Ericsson 2000: 210), their higher score might also hint at the fact that students respect different norms than professionals. After two years of intensive training, students might focus more on minor omissions or additions and source text equivalence than professional interpreters do after many years of experience and real life interpreting. A professional interpreter who is also a trainer mentioned during the retrospective interview that students might focus more on accuracy, while professionals are more aware of the importance of presentation and fluency in real life interpreting settings.

References

Cardoen, H., D'Amelio, N. (2013). The (inter)subjectivity of silent pauses in consecutive interpreting. In N. D'Amelio (ed.) *La recherche en interprétation: fondements scientifiques et illustrations méthodologiques*. Mons: CIPA, 121-137.

Dam, H. V. (1996). Text condensation in consecutive interpreting — summary of a PhD dissertation. *Hermes — Journal of Language and Communication in Business*, 17, 273-281.

Dam, H. V. (2007). What makes interpreters' notes efficient? Features of (non)efficiency in interpreters' notes for consecutive. In Y. Gambier, M. Schlesinger & R. Stolze (eds). *Doubts and directions in translation studies: selected contributions from the EST congress*. Amsterdam: John Benjamins, 183-198.

Ericsson, K. A. (2000). Expertise in interpreting: Insights from adopting an expert-performance perspective. *Interpreting 5*(2), 187-220.

Tang, F. & Dechao, L. (2017). A corpus-based investigation of explicitation patterns between professional and student interpreters in Chinese-English consecutive interpreting, *The Interpreter and Translator Trainer* 11(4), 373-395.

Matyssek, H. (1989). *Handbuch der Notizentechnik für Dolmetscher: ein Weg zur Sprachunabhängigen Notation.* Heidelberg: Julius Groos Verlag.

Rozan, J. (1956). *La prise de notes en interprétation consécutive*. Genève: Géorg.

Van Hoof, H. (1962). *Théorie et pratique de l'interprétation avec application particulière au français et à l'anglais*. Munich: Max Hueber.

Translation choices compared: Investigating translation variation in a learner translation corpus

Sara Castagnoli
University of Macerata
sara.castagnoli@unimc.it

This paper describes an empirical study aimed at investigating the concept of variation in translation through the analysis of different target versions (TTs) of the same source text (ST). More particularly, by observing the translation choices made by different translators with respect to specific points in the ST, the study seeks to identify which elements (e.g. lexical items, syntactic structures) are most open – and, conversely, less susceptible – to variation, as well as how different translation choices relate to the ideational, interpersonal and textual meaning (Halliday 1994) of the ST.

Previous research by Munday (2012) is taken as the main reference for this purpose. Analysing a corpus of two professional and 15 learner Spanish-English translations of a short story by Jorge Luis Borges, Munday found that variation was considerable on the paradigmatic axis, especially in the translation of elements with the highest interpretative and evaluative potential (e.g. adjective expressing attitude, in appraisal theory terms) and creative collocations; on the other hand, more invariance, or “stability”, was observed with respect to concrete nouns which do not have obvious competitor synonyms in the TL, determiners, personal pronouns/names, prepositions and other functional items. A second, comparable analysis of student translations in the technical domain revealed more frequent changes at the syntactic level than in literary texts, as well as surprising variation in the translation of technical terms, arguably the most significant words in the ST (i.e. it would be reasonable to expect content words which carry ideational meaning to be more stable in translation). Munday argues that translators’ (in)experience and domain-specific competence may play a major role in this case, although his analysis also confirms previous findings by Babych & Hartley (2004), who suggested that variation might be more frequent for words which have no obvious, ready-made translation equivalents, so that a stronger interpretative effort is required while translating; on the contrary, more stability would be observed in connection to TT words that are common dictionary equivalents or cognates to ST items. Munday also observed that invariance (i.e. word-forms that remain stable throughout all translations) was higher across the two professional translations (63-65% of the text) than in TTs commissioned to trainee translators (about 18%).

The present study sets out to assess whether Munday’s findings are corroborated by a comparable analysis of data extracted from a different learner translation corpus (LTC), here defined as a parallel corpus where several translations into the same target language, produced by translation trainees, are available for each source text. The corpus contains about 500 translations into Italian of several English and French STs, heterogeneous from the point of view of genre and topic. TTs were produced by Italian-native post-graduate students of specialised translation during ordinary examination sessions, which ensures that they are both ecologically valid (vs. elicited data) and comparable with respect to the conditions in which they were produced. While most LTC-based research has so far focused on the analysis of translation errors, the TTs in this corpus are not error-annotated, as it was mainly developed to enhance the identification of recurrent *features* of – rather than common *difficulties* in – learner translation (Castagnoli 2009). In order to judge to which extent translator behaviour may be affected by the specific genre of the ST, the analysis will be carried out on two sub-corpora representing different text types and degrees of specialization, including informative and persuasive texts in the field of economics and finance on the one hand, and more descriptive, evocative, literary extracts on the other.

Previous research based on this LTC, grounded in the so-called translation universals framework (Baker 1993), has focused primarily on variation – or lack thereof – in connective usage. Little variation was observed in the translations provided for specific connectives and their position within the sentence, with translators opting for one “preferred” option – usually a common dictionary equivalent for the ST connective – even when alternatives would have been possible (or even desirable, based on register considerations). Overall, interference from the

ST was identified as the overarching tendency, the reproduction of ST conjunctive patterns being far more frequent than both their normalization and explicitating/implicitating shifts (Castagnoli 2009, 2016). A second pilot investigation has focused on variation in the translation of different types of phraseological patterns, namely domain-specific compound terms, collocations/preferred combinations, and idioms. Although phraseology has often been investigated as an indicator of normalisation (see e.g. Kenny 1998; Dayrell 2007; Marco 2009), the tested hypothesis is that trainee translators may encounter difficulties in activating their phraseological competence in their own native language when translating out of an L2, because of the situation of language contact coupled to inexperience, thus producing more “strange strings” (Mauranen 2000). Preliminary analyses point to different degrees of variation across the three categories of multi-word units, possibly due to different causes – from the lack of knowledge of domain-specific terms to the non-existence of straightforward equivalent linguistic forms in the TL – and with different impacts on text meaning.

The present paper extends and re-interprets the above-mentioned findings, aiming in the first place at an improved description of variation based on the distinction between innocuous differences in lexicalization (“legitimate translation variation”, Babych & Hartley 2004), mistranslations, and other more or less acceptable choices affecting the text meaning to some extent. As in Munday’s study, STs and TTs will first be analysed manually in order to identify and classify types of (in)variance, then frequencies will be calculated to assess which items are more/less susceptible to it. In the second place, an attempt will be made to better characterise the relationship between invariance and ST interference by drawing on the concepts of “gravitational pull” (Halverson 2003, 2017) and “core patterns of lexical use” (Laviosa 1998). Although the study is essentially product-oriented and no triangulation with process-oriented data is possible, insights on the translation process might emerge by observing regularities and differences in translator behaviour.

References

Babych, B. & Hartley, A. (2004). Modelling legitimate translation variation for automatic evaluation of MT quality. *Proceedings of LREC 2004*, 833-836. Available online at http://www.mt-archive.info/LREC-2004-Babych-2.pdf.

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: John Benjamins, 233-250.

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

Castagnoli, S. (2016). Investigating trainee translators’ contrastive pragmalinguistic competence: a corpus-based analysis of interclausal linkage in learner translations. *The Interpreter and Translator Trainer* 3(10), 1-21.

Dayrell, C. (2007). A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics* 12(3), 375-414.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*, 2nd edition. London: Arnold.

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

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.* Berlin & Boston: De Gruyter & Mouton, 9-46.

Kenny, D. (1998). Creatures of habit? What translators usually do with words. *Meta* 43(4), 515-523.

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

Marco, J. (2009). Normalisation and the Translation of Phraseology in the COVALT Corpus. *Meta* 54(4), 842-856.

Mauranen, A. (2000). Strange Strings in Translated Language – A Study on Corpora. In M. Olohan (ed.) *Intercultural Faultlines*. Manchester: St. Jerome, 119-142.

Munday, J. (2012). *Evaluation in Translation: Critical Points of Translator Decision-Making*. London: Routledge.

Lewis Carroll's *The Queen of Hearts* in English and Czech: A corpus-stylistic approach to characterization through reporting verbs

Anna Čermáková, Michaela Mahlberg
University of Birmingham
a.cermakova@bham.ac.uk, m.a.mahlberg@bham.ac.uk

There is a growing amount of research on corpus stylistics and literary translation (e.g. Čermáková 2015; Mastropierro & Mahlberg 2017; Mastropierro 2017; Ruano 2017). One area where such research has proved to be particularly successful is the study of characterisation. An essential technique of characterisation is the way in which character speech is presented (e.g. Culpeper 2001). Speech presentation in narrative texts has therefore received a considerable amount of attention in the field of stylistics (e.g. Semino & Short 2004). Reporting verbs in particular are of interest from a corpus stylistics point of view as there is less variety in the reporting verbs than in the actual content of the speech that is being reported. Hence, it is possible to describe patterns of reporting verbs (e.g. Mahlberg et al. 2013). Corpus research into reporting verbs in Dickens's novels also suggests that the use of reporting verbs can be gendered, e.g. *pout* is mainly used with female characters whereas *growl* is used with male characters (Ruano forthcoming 2018).

At the same time, the frequent repetition of reporting verbs (specifically of the English *said*) constitutes a challenge for the translation into languages where repetition is not tolerated well, e.g. in translation from English to Czech (Levý 2011). But it also deserves attention more widely as the treatment of repetition as an undesirable stylistic feature is nearly "universal" (see e.g. Toury 1977, 1995; Ben-Ari 1998; Abdulla 2001). Recent corpus-based research into the translation of reporting verbs into Czech has confirmed that translators have a clear preference for variation and consequently for departure from the source text (Corness 2009; Farova 2016; Nadvornikova 2016).

Lewis Carroll's *Alice's Adventures in Wonderland* (first published 1865) and the subsequent *Through the Looking Glass* (1871) are not only two of the most analysed literary works but also two of the most translated ones (Horton 2002; Lindseth & Tannenbaum 2015). It is especially the verses, nonsense, wordplay and invented words that make the *Alice* text linguistically challenging for translators. However, other textual features identified by literary critics have received less attention. Feminist literary criticism, for instance, pointed out that the gender representation in the *Alice* books is rather unusual for Victorian literature. As Little (1976: 195) observed, the *Alice* books present "almost a comic compendium of feminist issues". Most of the novels' characters are male or are referred to as 'he'. There are only a few female characters: apart from Alice's cat Dinah, it is Alice herself, the Queen, the Duchess, her cook and Alice's sister. Literary opinion on these characters has been contradictory (see e.g. Garland 2008), especially concerning the main character Alice. All adult female characters are terrifying and the Queen stands in stark contrast to both Alice and the novel's weak male characters, particularly the King. This gender/power distribution is rather unusual for the time, even though parallels with Queen Victoria have been pointed out (Garland 2008: 29).

This paper aims to look at the iconic character of the Queen – in Carroll's own words "an embodiment of an ungovernable passion – a blind and aimless Fury" (Garland 2008: 28-29) in the English original and its most famous translation into Czech by Aloys Skoumal and Hana Skoumalová (first published in 1961). The paper analyses the reporting verbs that are used for the Queen and their Czech translation equivalents. For the purpose of this study, we focus on all reporting verbs that introduce speech, direct or indirect, and thought that appears in quotation marks. The study is based on a parallel aligned text of *Alice's Adventures in Wonderland* (available at www.korpus.cz, *InterCorp* corpus). We further use two reference corpora. For English, we use 19th century children's literature corpus (4.4 mil. words, available at http://clic.bham.ac.uk/) and for Czech we use a subcorpus of non-translated Czech children's fiction published between 1967 and 2013 (this may include republished texts with an earlier date of the first publication as is the case with *Alice*) (2 million words, created from SYN_v6 corpus available at www.korpus.cz). For the English source text the reference corpus reflects the time of its production while with the Czech translation we focus on the reception of the translation which – though dated – is the most recent and still widely read.

As we will show, the Queen is a loud, angry and terrifying character. The choice of reporting verbs is one of the key ways in which Carroll achieves this impression. There is only one characterizing adjective that Carroll uses to describe the Queen – *savage*. It has been claimed that "[i]n terms of traditional gender roles and language, the King occupies the feminine space while the Queen becomes the more dominant, masculine figure" (Garland 2008: 29). Our analysis supports this claim linguistically and shows how Carroll's choice of reporting verbs with the Queen, e.g. *shout* and *roar* which typically occur with male characters in the reference corpus, makes the Queen a masculine character. To support this picture of the Queen, we will further contrast the way her speech is presented with examples introducing the King's speech. We will further show that the picture is less clear in the Czech translation. The translation equivalents show a tendency for normalization as their choice is based on verbs that are less clearly associated with a character's gender. The patterns of reporting verbs that we discuss suggest a very subtle shift in the portrayal of the Queen for Czech readers. This shift, however, has implications for how the character of the Queen is situated within the Carroll scholarship.

References

Abdulla, A. K. (2001). Rhetorical repetition in literary translation. *Babel* 47(4), 289-303.

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

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 Linguistic* 20(3), 355-377.

Corness, P. (2009). Shifts in Czech translation of the reporting verb said in English fiction. In F. Čermák, P. Corness & A. Klégr (eds). *InterCorp: Exploring a Multilingual Corpus*. NLN: Praha, 159-176.

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

Fárová, L. (2016). Uvozovací slovesa v překladech třech různých jazyků. In A. Čermáková, L. Chlumská & M. Malá (eds). *Jazykové paralely*. Praha: NLN, 145-161.

Garland, C. (2008). Curious Appetites: Food, Desire, Gender and Subjectivity in Lewis Carroll's Alice Texts. *The Lion and the Unicorn* 32(1), 22-39.

Horton, D. (2002). Describing intercultural transfer in literary translation: Alice in Wunderland. In G. Thome, C. Giehl & H. Gerzymisch-Arbogast (eds). *Kultur und Übersetzung: Methodologische Probleme de Kulturtransfers*. Tübingen: Narr, 95-113.

Levý, J. (2011). *The Art of Translation*. (transl. by P. Corness, *Umění překladu* first published in Czech in 1963). Amsterdam: John Benjamins.

Lindseth, J. A. & Tannenbaum, A. (2015). *Alice in a World of Wonderlands: The Translations of Lewis Carroll's Masterpiece*. New Castle, DE: Oak Knoll Press.

Little, J. (1976). Liberated Alice: Dodgson's female hero as domestic rebel. *Women's Studies* 3(2), 195-205.

Mahlberg, M., Smith, C. & Preston, S. (2013). Phrases in Literary Contexts: Patterns and Distributions of Suspensions in Dickens's Novels. *International Journal of Corpus Linguistics* 18(1), 35-56.

Mastropierro, L. (2017). *Corpus Stylistics in Heart of Darkness and its Italian Translation*. London: Bloomsbury.

Mastropierro, L. & Mahlberg, M. (2017). Key words and translated cohesion — A corpus stylistic analysis of Lovecraft's At the Mountains of Madness and its Italian translation. *English Text Construction* 10(1), 78-105.

Nádvorníková, O. (2017). Les proportions des verbes SAY/DIRE/ŘÍCI dans les propositions incises et leurs équivalents en traduction: étude sur corpus parallèle. *Linguistica Pragensia* 27(2), 35-57.

Ruano, P. (2017). Corpus Methodologies in Literary Translation Studies: An Analysis of Speech Verbs in Four Spanish Translations of Hard Times. *Meta* LXII(1), 94-113.

Ruano, P. (forthcoming 2018). Charles Dicken's gender-based use of speech verbs: a stylistic analysis. *Gender and Language.*

Toury, G. (1977*). Translational Norms and Literary Translation into Hebrew, 1930-1945.* Tel Aviv: The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Comparing the incomparable? Rethinking n-grams for free word-order languages

Lucie Chlumská, David Lukeš
Charles University
lucie.chlumska@ff.cuni.cz, david.lukes@ff.cuni.cz

Ever since the boom of corpus-based resources, linguists have explored and employed various methods to identify and extract recurring language patterns from texts as these can reveal a lot about the syntagmatic nature of language and its grammatical, lexical and syntactic tendencies. One of these methods is the n-gram method based on extracting frequent sequences of n consecutive words. N-grams seem computationally and linguistically trivial but proved to be successful in identifying suitable candidate sequences of words worthy of further analytical attention. Like in many other linguistic areas, the majority of studies were carried out on the English language. N-grams in corpus linguistics were first used extensively by Biber et al. (1999) who identified a number of recurrent 4-6-grams occurring commonly in different register types.

Not much later, n-grams came into use in contrastive and translation corpus-based analysis. Baker (2004) tried various lengths of n-grams to compare translated and non-translated language, while in cross-linguistic contrastive studies, Forchini & Murphy (2008) analyzed 4-grams in Italian and English; Cortes (2008) analyzed 4-grams in English and Spanish; Ebeling & Ebeling (2013), in their book-length study, analyzed n-grams in English and Norwegian; Granger (2014) and Granger & Lefer (2013) used n-gram methodology in a comparison of English and French; and finally, Čermáková & Chlumská (2017) in a study of English and Czech place expressions.

As evidenced by the growing number of studies, the n-gram approach seems to be rather popular lately; however, it raises a number of serious methodological issues when applied cross-linguistically. One of the biggest challenges seems to be the issue of a suitable length of the n-gram as pointed out by Ebeling & Ebeling (2013), Granger (2014) or Čermáková & Chlumská (2017). The length of the n-gram may carry over in cross-linguistic analysis, but may also be substantially different (e.g. 4:4 as in EN: from side to side – CZ: *ze strany na stranu*, but also 4:1 as in EN: for the first time – CZ: *poprvé*). A major point raised by Granger (2014) is related to the contrastive study of typologically different languages, e.g. there may be variation in an n-gram in inflectional languages (EN: I am sure – CZ: *jsem si jistý/jistá*) that could possibly be resolved, as Granger suggests, by using lemmatization.

However, the rich variety of word forms belonging to one lexeme is not the only problem in such languages; the free word-order seems to be even more challenging as it strongly influences the very extraction of n-grams. As Čermáková & Chlumská (2017) report, if we look at the differences between more analytical English and highly inflectional Czech, there are approximately ten times more n-gram tokens above the same threshold found in English than in Czech. It is quite clear that the analytical nature of English relies to a much greater extent on patterning based on rigid sequences of words than free word-order and morphologically highly variable Czech. Patterning in an inflectional language is less regular and the patterns themselves allow for more extensive variability.

To provide an example, four words in the same structure, such as CZ: *myslel jsem si že* (EN: "I thought that"), can appear in several different combinations due to the free word order, and even not immediately next to one another: *jsem si myslel že* (EN: "I thought that"), *myslel jsem si původně že* (EN: "First I thought that") etc. Such differences are impossible to abstract over using n-grams, because these are always ordered and contiguous (although some positions within the n-gram may be underspecified, cf. skip-grams). Identifying such constructions is thus challenging, because none of the variants individually might make it above the given frequency cut-off point.

The adverse effects of free word order on n-gram frequency counts can be mitigated by breaking their rigid linear structure. One possible way of achieving this is by the following process:

1. slide a window of size n over the target corpus;
2. tally counts of all subsets of k < n elements taken from each window.

We call such subsets n-choose-k-grams, because they arise as the different k-element combinations over a given n-gram window. Obviously, the choice of n and k determines how computationally onerous the process will be. Compare this with n-grams: whereas with e.g. 4-grams, each 4-gram window in the corpus is tallied exactly once, with 8-choose-4-grams, each 8-gram window yields (8*7*6*5)/(4*5*3*2)=70 4-combinations.

These n-choose-k-grams have two desirable properties with respect to the task at hand:

1. being sets, k-combinations are unordered, i.e. word order differences are neutralized;
2. being subsets, k-combinations can abstract over extraneous words being inserted at any position within the original n-gram.

It can easily be seen how combinatorial explosion can make the task computationally intractable for improperly selected n and k (in particular, when n is large and k is about half). In practice, n and k should be selected in accordance with the locality principle: words that work towards a common goal may not always occur in the same order and may be interspersed with other words, but they will tend to occur in one another's close neighborhood. Sticking to this maxim should help yield combinations which span words which actually form a functional grouping, not just unrelated co-occurrences within a long n-gram window, as well as restrict n and k to manageable values.

Ironing out the details of the procedure is work in progress, but very early results are encouraging. In trying to identify the aforementioned CZ: *myslel jsem si že structure*, a 4-gram scan of our test corpus yields 4 different word order variants with frequencies 18, 17, 2 and 1 (one of which is spurious, an instance of another construction). By contrast, a 8-choose-4-gram scan yielded 177 matches within 8-gram windows, i.e. over three times as much. It remains to be determined whether this increase in recall is not outweighed by too great a decrease in precision.

References

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9(2), 167-193.

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

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.

Cortes, V. (2008). A Comparative Analysis of Lexical Bundles in Academic History Writing in English and Spanish. *Corpora* 3(1), 43-57.

Ebeling, J. & Ebeling, S. O. (2013). *Patterns in Contrast*. Amsterdam: John Benjamins.

Forchini, P. & Murphy, A. (2008). N-grams in comparable specialized corpora. Perspectives on phraseology, translation, and pedagogy. *International Journal of Corpus Linguistics* 13(3), 351-367.

Granger, S. (2014). A Lexical Bundle Approach to Comparing Languages: Stems in English and French. *Languages in Contrast* 14(1), 58-72.

Granger, S. & Lefer, M.-A. (2013). Enriching the phraseological coverage of high-frequency adverbs in English-French bilingual dictionaries. In K. Aijmer & B. Altenberg (eds). *Advances in Corpus-based Contrastive Linguistics: Studies in honour of Stig Johansson*. Amsterdam: John Benjamins, 157-176.

Coherence relations across speech and sign language: A comparable corpus study of additive connectives

Ludivine Crible[1], Sílvia Gabarró-López[2]
Université catholique de Louvain[1], Stockholm University[2]
ludivine.crible@uclouvain.be, silvia.gabarro@unamur.be

Discourse is built upon the connection between utterances through semantic-pragmatic relations such as addition, result or concession, among many others (e.g. Mann & Thompson 1988). These coherence relations are often signalled by connectives such as *and*, *so* or *but*. Sign languages, as natural languages, are no exception. In this paper, we contribute to the pioneering field of corpus-based discourse studies in sign languages (see also Gabarró-López 2017) by providing the first contrastive analysis of one type of coherence relation (viz. *addition*) and its connectives across a sign language (French Belgian Sign Language, henceforth LSFB) and a spoken language (French). LSFB is an under-studied minority language used in French-speaking Belgium by approximately 25,000 signers. LSFB has no written tradition, i.e. it is transmitted "orally" from one generation to another.

Connectives and the relations they signal vary across languages and registers, as previously shown in studies using either comparable (e.g. Kunz & Lapshinova-Koltunski 2015) or parallel corpora (e.g. Zufferey & Degand 2017). By contrast, very few studies have carried out multimodal comparisons of connectives across speech and writing (for English, see Biber et al. 1999; Fox Tree 2014), let alone across spoken and sign languages. Spoken corpora are increasingly used for quantitative (contrastive) discourse analyses (e.g. Taboada & Gómez González 2012 for English and Spanish; Kunz et al. 2017 for English and German; Crible 2018 for English and French), in spite of their relative rarity and smaller size compared to written corpora. Sign language corpora, on the other hand, have only emerged since the 2000s. Video recordings are highly costly to implement and to process: it is estimated that 1h of video data requires 250h for manual annotation. For this reason, corpus-based studies remain very scarce to date and tend to focus on other linguistic levels (phonology, morphology), leaving discourse analysis an open field to explore, especially from a contrastive perspective.

Within the wide panel of coherence relations, this study targets additive relations, also called "conjunction" in the Penn Discourse TreeBank (Prasad et al. 2008) or "Elaboration-additional" in Rhetorical Structure Theory (Mann & Thompson 1988). We chose the additive relation because of its very high frequency in corpus data: it is the most frequent relation in both written French (newspaper articles, Danlos et al. 2015) and spoken French (various registers such as conversations, interviews, news broadcast, Crible 2018). Our quantitative-qualitative analysis examines the number and types of connectives that can express an additive relation, in order to contrast its "markedness" (Asr & Demberg 2012) in the two languages, that is, whether *addition* is marked by dedicated connectives (e.g. English *in addition*) or by ambiguous, polyfunctional ones (e.g. English *and*).

To this end, we use two comparable samples of conversations (interactive dialogues, free exchange) taken from the Valibel corpus for spoken French (Dister et al. 2009) and the LSFB Corpus (Meurant 2015). All speakers and signers are Belgian and native.[1] The samples contain approximately 15,000 words (or signs) for each language. This relatively small corpus size is comparable to previous studies using spoken data (e.g. about 10,000 words in Taboada & Gómez-Gónzalez 2012) and is sufficient to obtain a substantial number of occurrences, given the high frequency of the additive relation. In fact, in the samples, we extracted 142 additive relations in French and 130 in LSFB.

[1] For the LSFB sample, there are two near-native signers; i.e. they have hearing parents (as 95% of deaf individuals) but they attended a boarding school for the deaf. They consider LSFB their first language.

We proceeded with a bottom-up methodology, by manually identifying any connective that expressed an addition between two utterances, as could be represented by the logical sign "Λ". This identification stage resulted in only two different connective types in spoken French, namely *et* 'and' (sometimes followed by *puis* 'then') and *en plus*, and four different signs in LSFB, namely PLUS (Figure 1), AND (Figure 2), ADD (Figure 3) and SAME (Figure 4). Our analysis shows that, while the number and frequency of additive connectives differ across French and LSFB, the two languages share a divide between dedicated (*en plus*, PLUS, AND, ADD) and polyfunctional connectives (*et*, SAME).

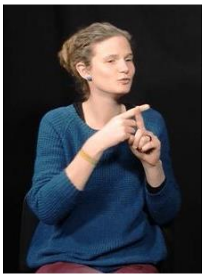


Figure 1. PLUS

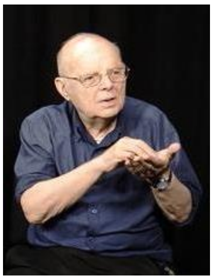


Figure 2. AND

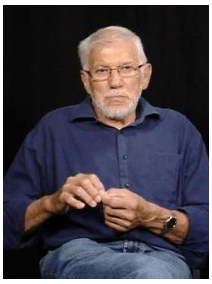


Figure 3. ADD

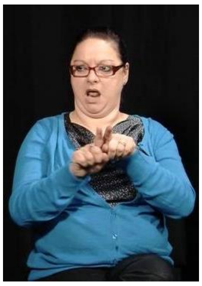


Figure 4. SAME

In a second step of the analysis, we focused on the two polyfunctional connectives *et* and SAME in order to investigate their other uses besides *addition*. To do so, we used a taxonomy of senses developed by Crible & Degand (2017), where 10 semantic labels are distinguished, including *addition* but also *contrast*, *consequence* or *alternative*. In addition, these labels can be further specified into one of four functional domains (*ideational*, *rhetorical*, *sequential*, *interpersonal*), depending on which aspect of the interaction the connective is targeting: connecting facts, reasoning arguments, turns, topics, opinions, etc. (cf. Halliday & Hasan 1976; Redeker 1991). Our analysis shows that, while *et* and SAME can both express other meanings besides *addition*, they differ in the extent of their functional spectrum as a result of their different semantic-pragmatic status: *et* in French is monosemous, that is, it only encodes addition even though it can be pragmatically enriched in context (for instance to express a *consequence*); by contrast, SAME in LSFB is polysemous and encodes both addition and comparison, which then leads to other meanings, such as *reformulation*.

In the paper, we report on the method and results of the contrastive study of additive connectives in French and LSFB, with a particular focus on the two highly polyfunctional connectives *et* 'and' and SAME. Quantitative and qualitative analyses will be combined to draw an exhaustive portrait of this basic relation of coherence in spoken and signed discourse. As the first of its kind, this new type of contrastive study has implications for the interpretation of sign languages, for the training of sign language interpreters and other professionals such as teachers working in bilingual LSFB-French settings, and for the education of deaf bilingual individuals.

References

Asr, F. & Demberg, V. (2012). Measuring the strength of linguistic cues for discourse relations. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA)*, Mumbai, India, 33-42.

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

Crible, L. (2018). *Discourse Markers and (Dis)fluency. Forms and Functions across Languages and Registers.* Amsterdam: John Benjamins.

Crible, L. & Degand, L. (2017). Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory* 13(1), 1-29.

Danlos, L., Colinet, M. & Steinlin, J. (2015). FDTB1, première étape du projet « French Discourse Treebank »: repérage des connecteurs de discours en corpus. *Discours* 17.

Dister, A., Francard, M., Hambye, P. & Simon, A.-C. (2009). Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de la banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de Linguistique* 33(2), 113-129.

Fox Tree, J.E. (2014). Discourse markers in writing. *Discourse Studies* 17(1), 64-82.

Gabarró-López, S. (2017). *Discourse markers in French Belgian Sign Language and Catalan Sign Language: BUOYS, PALM-UP and SAME*. Doctoral dissertation, Université de Namur.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski E., Menzel, K. & Steiner, E. (2017). GECCo — an empirically-based comparison of English-German cohesion. Empirical Translation Studies. In G. De Sutter, M.-A. Lefer & I. Delaere (eds). *New Methodological and Theoretical Traditions*. Berlin: Mouton de Gruyter, 265-312.

Kunz, K. & Laphinova-Koltunski, E. (2015). Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies* 14(1), 258-288.

Mann, W. & Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3), 243-281.

Meurant, L. (2015). *Corpus LSFB. Un corpus informatisé en libre accès de vidéos et d'annotations de la langue des signes de Belgique francophone (LSFB)*. Laboratoire de Langue des signes de Belgique francophone (LSFB-Lab), FRS-F.N.R.S et Université de Namur. www.corpus-lsfb.be.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 08)*, Marrakech, Morocco, 2961-2968.

Redeker, G. (1991). Linguistic markers of discourse structure. *Linguistics* 29, 1139-1172.

Taboada, M. & Gómez-Gónzalez, M. (2012). Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences* 6, 17-41.

Zufferey, S. & Degand, L. (2017). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory* 13(2), 1-24.

Parallel corpus of simultaneous interpreting as a tool of contrastive linguistics: Investigation of collocativity

Dayter Daria
University of Basel
daria.dayter@unibas.ch

The discipline of contrastive linguistics has experienced a rejuvenation with the introduction of large computerized multilingual corpora. Much research in the recent decades has demonstrated the usefulness of this tool in contrastive study of syntax (Ebeling 1998), pragmatics (Aijmer & Simon-Vandenbergen 2004) and other language levels. Translation corpora provide access to a much desired *tertium comparationis* for the comparison of two languages. As some researchers have argued, an even more promising approach is to combine the analysis of translation with the analysis of comparable data from the two languages (see Altenberg & Granger 2002).

The present project is situated in the area of translation and interpreting studies and investigates variation in original and simultaneous interpreted texts, i.e. attempts to approach the much-debated issue of "translationese" (or "interpretese"). For the purposes of the study, I am compiling a corpus of bidirectional Russian-English simultaneous interpreting (SIREN). Currently, SIREN contains approx. 229,000 words of political discourse (speeches, press-conferences, briefings) and consists of two subcorpora: Russian originals interpreted into English, and English originals interpreted into Russian. The material has been specifically chosen to include a large proportion of "free" simultaneous interpreting, which makes SIREN different from the other existing SI corpora which predominantly include interpreting "with text". The corpus includes a mixture of interpreting into the interpreters' A and B languages. The data is drawn from the UN Live Web TV, the video portal of the United Nations that broadcasts live and on demand UN meetings and events with parallel audio streams in the UN languages, as well as from televised broadcasts of interviews and press briefings of the Russian, US and UK governments that were made available to the researcher by the RT channel. The corpus is enriched with POS annotation, carried out with help of the CLAWS tagger for English and TreeTagger for Russian, and manual annotation for disfluencies. The parallel components of the corpus are time-aligned in five seconds intervals, which allows the researcher to work with a parallel concordance to compare the source and target texts if necessary.

In this study, I am concerned with the description of the simultaneously interpreted variety of English from the point of view of collocativity, and its comparison with non-translated English. Collocativity here is understood as the degree to which a register relies on recurrent word-combinations characteristic of the language in question. The assumption in early literature on translated varieties has been that translated language tends to be 'simplified' and 'normalised" (Baker 1995; Kenny 2001; Laviosa-Braithwaite 1996). Corpus-based studies focussing specifically on collocativity, although for translated rather than simultaneously interpreted texts, confirmed that normalization occurs in translation (Bernardini 2015). However, shallow statistical markers of SI material, such as lexical variety and density, show some tendencies in the opposite direction (Dayter 2018). The present study applies the methodology developed by Bernardini to SIREN to investigate whether simultaneously interpreted texts use more and/or stronger collocations than non-interpreted English. Part-of-speech collocational patterns are automatically extracted from SIREN and then rated according to the Mutual Information and logDice metrics based on ukWaC, a 2 billion word Web corpus. The use of ukWaC allows one to overcome the data bottleneck usually posed by small interpreting corpora that do not contain enough information to make reliable statistical claims. For the Russian subcorpus, the potential collocational patterns are first identified on the basis of the collocational dictionary of Russian, CrossLexica. The Russian National Corpus is used for rating the POS collocations extracted from SIREN.

Bernardini's (2015) findings unambiguously identified non-translated language as the register with higher collocativity. The results of the present analysis, however, show that the degree of collocativity in simultaneously interpreted language depends on the specific POS pattern. Significant differences for POS pattern lists ranked according to the strength of association and absolute frequency were found for six patterns. Only three of them evidence higher collocativity in original English, while the other three speak towards the fact that interpreted English makes greater use of more frequent and more strongly associated collocations.

One possible explanation is typological: with Russian marking tense, aspect and voice on the verb, translation into English results in higher numbers of auxiliary and light verbs that converge to form POS patterns. Another explanation relies on Shlesinger's (1989) hypothesis that simultaneous interpreting flattens the orality-literacy distinction between registers. This means that inherently oral genres, e.g. interviews, could become more *literate* when interpreted, while inherently literate genres, e.g. an opening speech delivered by the President of the General Assembly at a United Nations General Debate, could become more *oral*. A case can be made that higher collocativity is a marker of higher orality, since more varied, non-repeating vocabulary is seen as a literate feature. To fully explore this possibility, further investigation into the differences in collocativity between text types present in SIREN is necessary.

References

Aijmer, K. & Simon-Vandenbergen, A.-M. (2004). A model and a methodology for the study of pragmatic markers: the semantic field of expectation. *Journal of Pragmatics* 36, 1781-1805.

Altenberg, B. & Granger, S. (2002). Recent trends in cross-linguistic lexical studies. In B. Altenberg & S. Granger (eds). *Lexis in contrast*. Amsterdam: Benjamins, 3-48.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2), 223-243.

Bernardini, S. (2015). Translation. In D. Biber & R. Reppen (eds). *The Cambridge Handbook of Corpus Linguistics*, 515-536. Cambridge: Cambridge UP.

CrossLexica. https://www.gelbukh.com/xlex/.

Ebeling, J. (1998). Contrastive linguistics, translation, and parallel corpora. *Meta* 43(4), 602-615.

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

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

Laviosa-Braithwaite, S. (1996). *The English Comparable Corpus (ECC): a resource and a methodology for the empirical study of translation*. PhD Thesis, University of Manchester.

Shlesinger, M. (1989). *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum*. MA thesis, Tel Aviv University.

Meaning shifts in translation: A corpus-based Behavioral Profile approach

Pauline De Baets, Lore Vandevoordee, Gert De Sutter
Ghent University
Pauline.DeBaets@ugent.be, Lore.Vandevoorde@ugent.be, Gert.DeSutter@ugent.be

During the last two decades, research within corpus-based translation studies (CBTS) has been focusing on the question how translated texts differ from original, non-translated texts. A great number of studies have been dedicated to the different types of translational effects that are likely to occur in translated texts (De Sutter & Van de Velde 2010; Evert & Neumann 2017; Kruger 2012), with the bulk of these studies focusing on validating or refuting translational effects on the lexical, morphosyntactic and pragmatic level. As a result of this predominant research focus, numerous corpus-based studies carried out within the so-called universals paradigm have shown that language use in translations differs from language use in non-translated texts, for example that translations seem to conform more to the norms of the target language than non-translated texts (normalization, e.g. Delaere et al. 2012) or that translated texts tend to use more explicit expressions than original texts (explicitation, e.g. Olohan & Baker 2000) or that language use in translation is more oriented towards the source language (shining through, e.g. Teich 2003). The question, however, whether these specific features of translated language also occur on the semantic level, has rarely been asked. An example of such a semantic shift was found by Vandevoorde et al. (2017). According to their research, the subtle meaning differences between near-synonyms were mitigated in translated texts, which indicates that a certain semantic field in translated language is less differentiated than in non-translated language. If (subtle) lexicosemantic differences indeed appear to occur in translations compared to non-translated texts, that would undermine the core assumption of what translation defines, namely that there is semantic equivalence between source texts and their translations: “it seems to be firmly embedded in public opinion that in translation it is the meaning that has to remain unchanged” (Klaudy 2010:80). We will call this the semantic stability hypothesis, which states that the semantic structure of a given lexeme in translations is identical to its semantic structure in non-translated texts. Consequently, the main objective of this paper is to explore possible lexicosemantic translational effects and to verify the semantic stability hypothesis by investigating whether the (sub-)meanings associated with verbs of inchoativity shift during translation:

(a) based on the Firthean idea that the meaning of a word can be deduced from the context it keeps, we investigate the differences between the contextual properties of near-synonyms in translated and in non-translated language;
(b) we will calculate and compare the semantic distances between near-synonyms in the semantic field of inchoativity in translated and in non-translated language.

We compared the meaning structure of the field of inchoativity in a parallel corpus of English-to-Dutch translations to that of the same field in a comparable corpus of authentic Dutch texts. Both corpora are included in the Dutch Parallel Corpus (DPC), which is a multi-genre, sentence-aligned, 10-million-word corpus of Dutch, French and English (Macken et al. 2011). First, we had to select the lexemes that make up the semantic field of inchoativity. This was done by means of the semantic mirroring procedure (Vandevoorde et al. 2017), a method based on back-and-forth translation. The semantic mirroring procedure (SMM) enables us to look beyond the prototypical expressions found in the dictionary and consequently to create a broader semantic field. The SMM provided us with 5 lexemes of inchoativity (*beginnen* [to begin], *starten* [to start], *van start gaan* [to take off], *opstarten* [to start up] and *aanvatten* [to commence]). Secondly, all the sentences containing one of the 5 lexemes under study are extracted from the corpus and are annotated for different contextual parameters. We did not focus on any predefined meaning of our lexemes under study, but instead we used contextual information to uncover possible syntactic, semantic or pragmatic differences between the near synonyms of our semantic field. To do so, we applied the Behavioral Profile method (Gries & Divjak 2009; Szymor 2015), “a usage-based method that aims at capturing the complexity of word meaning by looking at the contextual features of the words under study” (Szymor 2015: 486). We opted for this particular approach because it is an objective, precise and corpus-based method that proves to be suitable to capture the complexity of word meaning and that allows for bottom-up identification of distinctive features (Szymor 2015). Behavioral profiling provides co-occurrence data of many different kinds, as it focuses on every property of the linguistic context of a lexeme under study. Those properties are called ID-tags and explore the entire semantic, morphological and syntactic context of the retrieved lexeme. In particular, we manually annotated 665 sentences for 31 ID-tags, such as animacy and concreteness of the subject and object, semantic category of the subject and object, mode of the verb, object type, presence of a modifying or modified verb, etc. Because these ID-tags represent the syntactic and semantic architecture of an individual lexeme, this annotation enabled us to explore in which respect the lexeme is unique and whether the profile of a lexeme remains stable in translational data compared to authentic, non-translated data. In a final step, the enriched corpus data were statistically analyzed. By means of multivariate, statistical procedures (viz. Random Forests and correspondence regression), we visualized and subsequently interpreted the structure of the semantic field of inchoativity in translated and non-translated language. The results of our analyses show that the behavioral profiles of the lexemes under study do not remain stable across the different corpus components, running counter to the posited "semantic stability hypothesis". Indeed, the discriminant ID-tags in translated language differ from those in non-translated language, pointing at a slightly different organization of the semantic field of inchoativity in translated language, compared to non-translated language. Hence, the semantic stability hypothesis is refuted. These findings do not only help to fill the 'semantic' research gap within corpus-based translation studies, but they can also affect the field of contrastive linguistics, as they can undermine one of the core assumptions underlying the use of parallel corpora in contrastive linguistics, namely that there is (perfect) semantic equivalence between source texts and their translations.

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: John Benjamins, 233-250.

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

De Sutter, G. & Van de Velde, M. (2010). Determinants of syntactic variation in original and translated language: a corpus-based study of PP placement in German. *International Journal of Translation*, 22(1), 59-76.

Evert, S. & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In G. De Sutter, M.-A. Lefer, I. Delaere (eds). *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin: De Gruyter, 47-80.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*. Oxford: Blackwell, 1-32.

Gries, S. T. & Divjak, D. (2009). Behavioral profiles: a corpus-based approach to cognitive semantic analysis. *New directions in cognitive linguistics*, 57-75.

Klaudy, K. (2010). Specification and Generalisation of Meaning in Translation. *Meaning in translation*, 19, 81-103.

Kruger, H. (2012). A corpus-based study of the mediation effect in translated and edited language. *Target,* 24(2), 355-388.

Macken, L., De Clercq, O. & Paulussen, H. (2011). Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta* 56(2), 374-390.

Olohan, M. & Baker, M. (2000). Reporting that in translated English. Evidence for subconscious processes of explicitation? *Across languages and cultures*, 1(2), 141-158.

Szymor, N. (2015). Behavioral Profiling in Translation Studies. *trans-kom*, 8(2), 483-498.

Vandevoorde, L., Lefever, E., Plevoets, K. & De Sutter, G. (2017). A corpus-based study of semantic differences in translation : the case of inchoativity in Dutch. *Target*, 29(3), 388-415.

Teich, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts* (Vol. 5). Berlin: Walter de Gruyter.

Positive evaluative adjectives in promotion-oriented texts: An exploratory study using a translation corpus

Sylvie De Cock
Université catholique de Louvain, Université Saint-Louis Bruxelles
sylvie.decock@uclouvain.be

According to Hunston (2011: 1), evaluative language is 'language which indexes the act of evaluation or the act of stance-taking (…) and expresses an attitude towards a person, situation or other entity'. Research has shown that evaluative language is prevalent in promotional discourse (advertising is a classic and prototypical example of promotional discourse, Bhatia 2005) and in promotion-oriented texts that present things, companies or cities in a favourable light (e.g. Kranich 2016; Ho & Suen 2017). Evaluative language is interactive in nature (Partington et al. 2013) as it creates 'an ideology shared by writer and reader' (Hunston 2011: 12), which arguably ties in well with the persuasive character of promotion-oriented discourse.

As pointed out by Kranich (2016: 75), '[a]djectives represent the word class most closely associated with evaluation' (see also Maat 2007). Adjectives like 'wonderful' can even be labelled as prototypically evaluative as their overt and only purpose is to evaluate (Channel 1999). In Partington et al.'s words, the 'evaluative weight' of prototypically evaluative adjectives is 'intrinsic' or 'in-built' and evaluation is 'a major if not predominant part of their function' (2013: 52). According to Kranich (2016: 76), the syntactic use of evaluative adjectives (predicative or attributive) is well worth studying because of the effect that attributive use can have on the message: 'the use of adjectives in attributive position lends the evaluation a more objective flavor than the use of evaluative adjectives in predicative position.'

This study sets out to present the results of an exploratory study into the use and translation of positive evaluative adjectives in the 'Label France' corpus (Centre for English Corpus Linguistics, Université catholique de Louvain). The 'Label France' corpus is a unidirectional translation corpus (source language: French; target language: English) made up of articles from the quarterly magazine entitled 'Label France', published by the French Ministry of Foreign Affairs. The corpus contains over 800,000 words of informational texts that portray France (tourism, culture, economy, etc.) in a positive way in each language.

Although not primarily promotional in nature, the texts included in the corpus, like many informational texts these days (Bhatia 2005), also incorporate promotional elements. The focus of this study is positive evaluative adjectives (e.g. 'excellent', 'prestigieux', 'exceptionnel', 'formidable', 'remarquable' or 'magnifique' in the original texts in French). The adjectives selected for inclusion in the analysis are all prototypically evaluative and display positive evaluation. They were identified on the basis of a frequency list of word forms from the French source language texts using Wordlist (WordSmith Tools) and careful examination in context. Some French adjectives like 'grand' were excluded from the analysis because, although they can be used to denote positive evaluation as in (1), they tend to be used prototypically to indicate size (cf. (2)). A frequency list of word forms from the English target language texts was also used to identify the positive evaluative adjectives in the translated texts that do not have a corresponding adjective equivalent in French (e.g. 'successful').

(1) Pour la communauté scientifique internationale, ce fut un grand savant. Pour le public, un humaniste généreux. (French = source language; 'a great scholar')

(2) Qui dit cheval sur grand écran dit souvent western ou film d'aventures – et pense cinéma américain (French = source language; 'big screen')

The study aims to answer the following research questions:

- To what extent are the positive evaluative adjectives in the source texts translated as positive evaluative adjectives in the target texts? To what extent are the positive evaluative adjectives in the translated texts

translations of positive evaluative adjectives in the source texts? To what extent do the translated items exhibit a similar degree of positive evaluation?
- To what extent are the positive evaluative adjectives in the source texts and their translations in the target texts used in attributive position?
- What are the preferred collocational patterns of the positive evaluative adjectives in the source texts and of their translations in the target texts?

More generally, this exploratory investigation also examines which elements tend to be portrayed in a positive light using positive evaluative adjectives in the source texts (and their translations in the target texts).

Preliminary findings suggest that the English translations of the French positive evaluative adjectives under study also exhibit positive flavours, which goes some way towards preserving the promotion-oriented character of the texts. It is interesting that in some cases the positive evaluation is even given an extra boost in the translation. This is illustrated by the use of the superlative form of the adjectives in the English translation in (3).

(3) Antoine de Saint-Exupéry, qui avait survolé tant de beaux et grandioses paysages à travers le monde, écrivait au printemps 1944, au retour d'une mission qui précéda de peu son dernier voyage (French = source language)

Antoine de Saint-Exupéry, who had flown over so many of the world's **most** beautiful and magnificent landscapes, wrote in Spring 1944, on his return... (English = target language)

It also emerges from the analysis that the positive evaluative adjectives in the source and target texts are overwhelmingly used in attributive position, which is not unexpected given Kranich's (2016) observation (cf. above).

References

Bhatia, V. K. (2005). Generic patterns in promotional discourse. In H. Halmari & T. Virtanen (eds). *Persuasion across Genres*. Amsterdam & Philadelphia: John Benjamins Publishing Company, 213-225.

Channel, J. (1999). Corpus-based analysis of evaluative lexis. In S. Hunston & G. Thompson (eds). *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 38-55.

Ho, V. & Suen, A. (2017). Promoting a city's core values using evaluative language. *International Journal of Applied Linguistics* 27(1), 286-308.

Hunston, S. (2011). *Corpus Approaches to Evaluation. Phraseology and Evaluative Language*. New York: Routledge.

Kranich, S. (2016). *Contrastive Pragmatics and Translation. Evaluation, epistemic modality and communicative styles in English and German*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Maat, H. P. (2007). How promotional language in press releases is dealt with by journalists. Genre mixing or genre conflict? *Journal of Business Communication* 44, 59-95.

Partington, A., Duguid, A., Taylor, C. (2013). *Patterns and Meanings in Discourse. Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Expressing stance in parliamentary debates: A French-Spanish corpus-driven study

Barbara De Cock, Marie-Aude Lefer
Université catholique de Louvain
barbara.decock@uclouvain.be, marie-aude.lefer@uclouvain.be

This presentation reports on an exploratory, corpus-driven study of French and Spanish stance expressions in debates held at the European Parliament. Stance is here understood as “the speaker’s evaluative, epistemic or affective positioning towards a stance object” (Kärkkäinen 2012: 2195). More specifically, our paper focuses on recurrent multi-word stance expressions, such as French *à mon avis* ‘in my opinion’, *il me semble* + ADJ + *de* ‘it seems + ADJ + to’, *il est essentiel de* ‘it is essential to’ and Spanish *yo creo que* ‘I believe that’, *me gustaría* ‘I would like to’, *es necesario* ‘it is necessary’, and takes stocks of insights from both contrastive pragmatics (Aijmer 2011) and contrastive phraseology (Ebeling & Oksefjell Ebeling 2013; Granger 2014).

Our study relies on comparable Europarl subcorpora of European parliamentary proceedings in French and Spanish (see Koehn 2005), corresponding to approximately 10 years of debates (up to 2010) and 3 million tokens per language. The subcorpora used in this study are restricted to verbatim reports of speeches originally delivered in French and Spanish by Members of Parliament (see Cartoni & Meyer’s directional version of Europarl, which relies on Europarl’s language tag). In other words, they do not contain any translated texts. Even though it must be acknowledged that French and Spanish are occasionally used by non-native speakers in the European Parliament, in the vast majority of cases, the two languages are used by Members of Parliament from France, French-speaking Belgium and Spain, respectively.

In this paper, we adopt a corpus-driven approach to stance. The stance data were obtained by automatically extracting n-grams (also called ‘lexical bundles’), which are recurrent sequences of *n* contiguous words, i.e. “sequences of word forms that commonly go together in natural discourse” (Biber et al. 1999: 90). As pointed out by Granger (2014: 69), n-grams “are a powerful window onto pragmatics and rhetoric. It is undeniably a quick-and-dirty method, but one that has great heuristic power: it generates a multitude of word sequences that have so far received very little interest in the contrastive literature”. In this study, we have extracted 2- to 5-grams with a minimum frequency of 50 occurrences per million words. The automatically retrieved n-grams were then manually analyzed in context so as to identify stance expressions and patterns. Methodologically, the comparative analysis of French and Spanish, which are arguably quite close, being two Romance languages, raises a number of issues. One of them is that Spanish is a pro-drop language, while French is not. As a result, Spanish stance expressions containing a conjugated verb form may either be 1- or 2-grams, depending on the optional overt expression of the subject (e.g. (*yo*) *creo* ‘I believe’, (*nosotros*) *pensamos* ‘we think’), while in French, corresponding expressions typically contain two words (*je crois* ‘I believe’, *nous pensons* ‘we think’). In order to ensure an optimal cross-linguistic comparability of the datasets analyzed, a number of additional single-word stance expressions corresponding to the multi-word sequences identified through the n-gram approach were also extracted (e.g. *lamento/je regrette* ‘I regret’, *desearía/je souhaiterais* ‘I would like to’).

The presentation of the results will proceed in three steps. First, we will provide a structured, contrastive overview of the hundreds of stance expressions uncovered through the n-gram approach adopted in this paper, paying special attention to well-known French-Spanish morphosyntactic contrasts, such as the compulsory subject expression in French vs. the pro-drop character of Spanish, and the wider use of verbs with an experiencer dative in Spanish (e.g. *me gustaría decir* ‘I would like to say’; cf. Vázquez Rozas 2016). Second, we will zoom in on the different uses of formally similar or cognate forms, such as *croire/creer* ‘believe’ and *penser/pensar* ‘think’. Indeed, as repeatedly shown in studies on French and English political interviews (Fetzer & Johansson 2010), English and Spanish parliamentary enquiries (Marín Arrese 2015) and Catalan and Spanish parliamentary debates (De Cock & Nogué Serrano 2017), the use of cognate forms and seemingly similar patterns may in fact differ significantly across languages, mainly in terms of frequency, distribution and pragmatic/argumentative functioning. Finally, as noted by Goethals & Blancke (2014) in relation to thanking in

French, Spanish and Dutch in the European Parliament, speakers of different languages adhere to different discursive conventions, even within the same parliament. In our paper, we will examine if and to what extent multi-word stance expressions can be used to uncover the variation of discursive conventions concerning stance-taking across linguistic communities.

In our conclusion, we will sketch our future work, which mainly consists in extending the analysis to other languages (Dutch and English) and additional registers (newspaper editorials and research articles; cf. Neumann 2013, 2014) and in using parallel corpus data to analyze the impact of translation on some of the stance-related discursive and pragmatic traits of parliamentary proceedings.

References

Aijmer, K. (ed.) (2011). *Contrastive Pragmatics*. Amsterdam & Philadelphia: Benjamins.

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

Cartoni, B. & Meyer, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *8th International Conference on Language Resources and Evaluation* (LREC).

De Cock, B. & Nogué Serrano, N. (2017). The pragmatics of person reference: A comparative study of Catalan and Spanish parliamentary discourse. *Languages in Contrast* 17(1), 96-127.

Ebeling, J. & Oksefjell Ebeling, S. (2013). *Patterns in Contrast*. Amsterdam & Philadelphia: Benjamins.

Fetzer, A. & Johansson, M. (2010). Cognitive Verbs in Context. A Constrastive Analysis of English and French Argumentative Discourse. In S. Marzo, C. Heylen & G. De Sutter (eds). *Corpus Studies in Contrastive Linguistics*. Amsterdam & Philadelphia: Benjamins, 240-266.

Goethals, P. & Blancke, B. (2013). Un estudio exploratorio de las convenciones discursivas en el parlamento europeo: los agradecimientos en español, francés y neerlandés. *Revista de Lingüística y Lenguas Aplicadas* 8, 171-185.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. In M.-A. Lefer & S. Vogeleer (eds). *Genre- and Register-related Discourse Features in Contrast*. Special issue of *Languages in Contrast* 14(1), 58-72.

Kärkkäinen, E. (2012). *I thought it was very interesting*. Conversational formats for taking a stance. *Journal of Pragmatics* 44, 2194-2210.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit X*, 79-86.

Marín Arrese, J. I. (2015). Epistemic legitimization and inter/subjectivity in the discourse of parliamentary and public inquiries. A contrastive study. *Critical Discourse Studies* 12(3), 261-278.

Neumann, S. (2013). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: De Gruyter Mouton.

Neumann, S. (2014). Cross-linguistic register studies: Theoretical and methodological considerations. In M.-A. Lefer & S. Vogeleer (eds). *Genre- and Register-related Discourse Features in Contrast*. Special issue of *Languages in Contrast* 14(1), 35-57.

Vázquez Rozas, V. (2006). *Gustar*-Type Verbs. In J. Clancy Clements & J. Yoon (eds). *Functional Approaches to Spanish Syntax. Lexical Semantics, Discourse and Transitivity*. Basingstoke: Palgrave Macmillan, 80-114.

Explicitation and implicitation of Dutch and German noun-noun compounds in translation

Hinde De Metsenaere
Ghent University
Hinde.DeMetsenaere@UGent.be

Explicitation and implicitation are two translation studies concepts that have given rise to a vast array of studies (cf. Becher 2011: 20-76 for an overview). These studies are, however, of a very heterogeneous nature. This heterogeneity makes it difficult, if not impossible to compare the findings of these studies and to come to conclusive insights into the meaning of explicitation and implicitation in and for translation and translation studies (cf. Kamenická 2007: 45, Becher 2010: 3-4, Murtisari 2013: 315). The reason is that explicitation and implicitation have been interpreted differently, not rarely intuitively, by many translation studies researchers. This is due to the fact that the underlying concepts of explicitness and implicitness were not satisfactorily defined for translation studies purposes for a long period of time (cf. Murtisari 2013: 315).

Recently, De Metsenaere & Vandepitte (2017) have approached explicitation and implicitation from a relevance-theoretic perspective. The main assumption underlying this approach is that translation is an act of communication, in which utterances rather than sentences are translated from one language into another. Sentences are linguistically encoded constructs that can have many different meanings in as many contexts, whereas utterances are always part of a bigger unit that will limit their meaning possibilities. Relevance theory – already introduced into translation studies in 1991 by Gutt and further employed in Alves & Conçalves (2003) – is considered the most appropriate theoretical framework for explicitation and implicitation, because of its cognitive and pragmatic approach on the one hand, and its explicit-implicit distinction on the other hand, which directly lends itself to defining explicitness and implicitness. Based on relevance theory, De Metsenaere & Vandepitte (2017) have redefined not only explicitness and implicitness, but also explicitation, implicitation and three other notions that explicitation and implicitation need to be distinguished from when doing corpus-based research: addition, omission and substitution. These definitions are the starting point for the corpus-based study that is discussed in this paper.

Focus of the study is on Dutch and German nominal compounds with a nominal first constituent. These noun-noun compounds constitute a very productive word formation category in both languages, firstly because they are compounds with the smallest number of morphological constraints and the largest capacity for recursion, i.e. building compounds whose constituents are compounds themselves (on Dutch: Booij 1992: 37f, De Haas & Trommelen 1993: 375, Booij & Van Santen 1998: 148f; on German: Ortner et al. 1991: 9ff, Becker 1992: 7ff, Donalies 2005: 60ff, Eisenberg 2013: 217; on both: Hüning & Schlücker 2010: 791). Therefore, noun-noun compounds are often considered to be prototypical of nominal compounds in general. Secondly, noun-noun compounds have a meaning potential that is considerably high. Their meaning goes beyond the sum of the meanings of their constituents. This can be illustrated by Heringer's legendary example *Fischfrau* (fish-woman), which can be allotted more than ten different interpretations, from 'woman who sells fish' to 'woman who descends from a fish' (Heringer 1984: 2). The example shows that all interpretations must involve some relation between the constituents – here, between a woman and a fish –, but the exact nature of that relation is left unspecified (on Dutch: De Caluwe 1991: 131, Booij & Van Santen 1998: 153; on German: Heringer 1984: 3, Schlücker 2012: 14, Eisenberg 2013: 220). This does not mean, however, that noun-noun compounds are cryptic semantic constructs. On the contrary, they offer the relevant amount of information that is needed to guide the language user to the intended interpretation. The meanings of the constituents, the information that the second constituent is modified by the first, and above all the context in which the compound is introduced are sufficient to infer the tacit relation between the constituents and, hence, the meaning of the compound (on Dutch: De Caluwe 1991: 131, Booij & Van Santen 1998: 147f; on German: Heringer 1984: 5ff, Mohammed 2011: 71, Schlücker 2012: 15, Eisenberg 2013: 220f). Although nominal compounding is a very productive word formation category in Dutch and German, it is often claimed that it is realized differently in these closely related languages. Where German prefers a compound, Dutch may opt for an alternative construction (Booij & Van Santen 1998: 148; Campe 2010: 208; Hüning & Schlücker 2010: 791ff; Hüning 2010) that can, but must not necessarily, lead to differences in meaning.

The study that is reported on in this paper examines explicitation, implicitation (and necessarily also addition, omission and substitution) of the meaning of Dutch and German noun-noun compounds in translation. The PAND corpus (*Parallelkorpus Niederlandisch-Deutsch*) is used, which is an electronic, bidirectional, two-million-word translation corpus of Dutch and German fiction and non-fiction texts that is compiled at the Ghent University Department of Translation, Interpreting and Communication. From this corpus, a balanced set of 10,000 noun-noun compounds is manually extracted. Through meaning analysis, the hypothesised asymmetry between explicitation and implicitation in translation (Klaudy 2001 in Klaudy & Károly 2005: 14) is addressed. It is examined if such an asymmetry can indeed be observed for the translation pair Dutch-German. Furthermore, the questions are answered if and to what extent explicitation and implicitation relate to translation direction, text genre and information distribution.

References

Alves, F. & Gonçalves, J. L. (2003). A Relevance Theory approach to the investigation of inferential processes in translation. In F. Alves (ed.) *Triangulating translation: perspectives in process oriented research*. Amsterdam & Philadelphia: John Benjamins, 3-24.

Becher, V. (2010). Towards a More Rigorous Treatment of the Explicitation Hypothesis in Translation Studies. *trans-kom* 3(1), 1-25.

Becher, V. (2011). *Explicitation and Implicitation in Translation Studies. A Corpus-Based Study of English-German and German-English translations of business texts.* PhD thesis, Universität Hamburg.

Becker, T. (1992). Compound in German. *Rivista di Linguistica,* 4(1), 5-36.

Booij, G. (1992). Compounding in Dutch. *Rivista di Linguistica,* 4(1), 37-59.

Booij, G. & Van Santen, A. (1998). *Morfologie. De woordstructuur van het Nederlands* (2nd ed.). Amsterdam: Amsterdam University Press.

Campe, P. (2010). Syntactic variation in German adnominal constructions: an application to the alternatives 'genitive', 'apposition' and 'compound'. In A. Lenz & A. Plewnia (eds). *Grammar between norm and variation*. Frankfurt am Main: Peter Lang, 193-218.

De Caluwe, J. (1991). *Nederlandse nominale composita in functionalistisch perspectief*. 's-Gravenhage: SDU Uitgeverij.

De Haas, W. & Trommelen, M. (1993). *Morfologisch handboek van het Nederlands*. Leiden: Instituut voor Nederlandse Lexicologie.

De Metsenaere & H. & Vandepitte, S. (2017). Towards a Theoretical Foundation for Explicitation and Implicitation. *trans-kom,* 10(3), 385-419.

Donalies, E. (2005). *Die Wortbildung des Deutschen. Ein Überblick* (2nd ed.). Tübingen: Gunter Narr Verlag.

Eisenberg, P. (2013). *Grundriss der deutschen Grammatik. Band eins: Das Wort* (4th ed.). Stuttgart & Weimar: J.B. Metzler.

Gutt, E.-A. (1991). *Relevance and Translation*. Manchester: St. Jerome Publishing.

Heringer, H. J. (1984). Wortbildung: Sinn aus dem Chaos. *Deutsche Sprache,* 12, 1-13.

Hüning, M. (2010). Adjective + Noun Constructions Between Syntax and Word Formation in Dutch and German. In A. Onysko & M. Sascha (eds). *Cognitive Perspectives on Word Formation*. Berlin, New York: Mouton de Gruyter, 195-218.

Hüning, M. & Schlücker, B. (2010). Konvergenz und Divergenz in der Wortbildung — Komposition im Niederländischen und im Deutschen. In A. Dammel, S. Kürschner & D. Nübling (eds). *Kontrastive Germanistische Linguistik*. Hildesheim & Zürich & New York: Georg Olms Verlag, 783-825.

Kamenická, R. (2007). Defining explicitation in translation *Brno Studies in English* (Vol. 33). Brno: Masarykova univerzita, Filozofická fakulta, 45-57.

Klaudy, K. (2001). *The Asymmetry Hypothesis. Testing the asymmetric relationship between explicitations and implicitations.* Paper presented at the Third International Congress of the European Society for Translation Studies: Claims, Changes and Challenges in Translation Studies.

Klaudy, K. & Károly, K. (2005). Implicitation in translation: Empirical evidence for operational asymmetry in translation. *Across Languages and Cultures,* 6(1), 13-28.

Mohamed, N. E. A. (2011). Deutsche Nominalkomposita und ihre Übersetzungsproblematik ins Arabische. *Lebende Sprachen,* 56(1), 65-76.

Murtisari, E. T. (2013). A Relevance-based Framework for Explicitation and Implicitation in Translation. An Alternative Typology. *trans-kom,* 6(2), 315-344.

Orntner, L., Müller-Bollhagen, E., Ortner, H., Wellmann, H., Pümpel-Mader, M. & Gärtner, H. (1991). *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Vierter Hauptteil: Substantivkomposita.* Berlin & New York: Walter de Gruyter.

Schlücker, B. (2012). Die deutsche Kompositionsfreudigkeit. Übersicht und Einführung. In L. Gaeta & B. Schlücker (eds). *Das Deutsche als kompositionsfreudige Sprache*. Berlin & Boston: Walter de Gruyter, 1-25.

Disentangling the motivations underlying syntactic explicitation in contact varieties: A MuPDAR analysis of *that* vs. zero complementation

Gert De Sutter[1], Haidee Kruger[2,3]
Ghent University[1], Macquarie University[2], North-West University[3]
gert.desutter@ugent.be, haidee.kruger@mq.edu.au

Corpus-based translation studies (CBTS) have seen many significant developments in recent years, including the compilation of multimodal corpora (e.g. Alves et al. 2010; Serbina et al. 2015; Bernardini et al. 2016); the adoption of advanced statistical techniques to characterize translational choices more accurately (e.g. Oakes & Ji 2012; De Sutter et al. 2017); triangulating product, process and experimental data to get a better insight into the underlying processing mechanisms underlying translational output (Hansen-Schirra 2017); the analysis of intermediate versions of translated texts and the influence of (post-)editorial interventions (e.g. Kruger 2012; Bisiada 2016); and – albeit somewhat more modestly – theoretical innovation (e.g. Halverson 2017).

What has greatly contributed to these developments is the *de-isolation* of CBTS. Whereas in its early years, CBTS as scientific endeavor strongly focused on the unique characteristics of translation with the aim of setting it apart from other disciplines (especially contrastive linguistics), more recently CBTS has witnessed a process of tearing down the interdisciplinary walls (e.g. Kruger & Van Rooy 2016). The insights and methodologies of neighboring disciplines – a.o. interpreting studies, contrastive linguistics, SLA, variational linguistics, sociology, psycholinguistics, contact linguistics – have greatly contributed to the recent descriptive, analytical and theoretical successes in CBTS. As a consequence of these developments, CBTS is gradually dissolving in the broader discipline of empirical translation studies – corpora being only one of the methodologies being deployed.

The present paper aims to contribute to these developments by re-analyzing a classic topic in CBTS, viz. the syntactic alternation between *that* and zero in English complement clauses (Olohan & Baker 2000; cf. example 1), within the context of translation as a contact variety.

(1) Adi Goldberg, a 17-year-old high school student, confessed {**that / Ø**} she was scared when she heard the explosion (dpc-ind-001800-en).

In particular, this study analyzes how this alternation is used in central (non-contact) and 'peripheral' (contact) varieties of English by focusing on the language-internal and -external factors that influence the choice between one of the syntactic alternatives. The aim is to contribute to both theoretical and methodological refinement in this area of research. The theoretical contribution centers on disentangling the motivations that have been proposed for the increased explicitness of translated language, namely cognitive effort or complexity, cross-linguistic influence, and risk avoidance (see Kruger to appear). We do so against the background of proposals that increased explicitness is not only a feature of translated language, but of other language-contact varieties as well, such as L2 varieties.

For this study, we consider British English as the central variety, reflected in a register-differentiated corpus compiled from the International Corpus of English for Great Britain (ICE-GB)[1] and the British English texts in the Dutch Parallel Corpus (Macken et al. 2011). For the contact varieties, we use register-differentiated corpora that reflect two types of contact varieties: translations and highly proficient L2 writing. For each category, we include register-differentiated corpora with Afrikaans and Dutch as source/first languages. While the two languages are closely related, they demonstrate distinct preferences for complementizer omission (with Afrikaans much more permissive than its Dutch parent in complementizer omission).

[1] See http://ice-corpora.net/ice/.

- **Translation**: Translations from Dutch are compiled from the Dutch Parallel Corpus (DPC), while translations from Afrikaans are from a self-compiled corpus.
- **Proficient L2 writing**: To match the translation corpora, we use two self-compiled corpora of written L2 English (with Dutch and Afrikaans as first languages).

In the South African context, native South African English (rather than British English) may function as local reference variety, and has also been shaped by long-term contact with Afrikaans. To account for this potential contact effect with Afrikaans, we also include a corpus of native South African English, using written components of the International Corpus of English for South Africa (ICE-SA), currently under construction.

An overview of all the corpora used can be found in the table below:

Corpus (component)	Registers	Total word count
Original British English texts (DPC)	Journalistic texts, manuals, political speech, corporate communication, tourism, press releases, novels, nonfiction	1,000,000
Original British English texts (ICE)	Academic, fiction, instructional, persuasive, popular, reportage	250,000
Original South African English texts (ICE)	Academic, fiction, instructional , persuasive, popular, reportage	180,000
Translated English texts from Dutch (DPC)	Journalistic texts, manuals, political speech, corporate communication, tourism, press releases, novels, nonfiction	1,000,000
Translated English texts from Afrikaans	Academic, fiction, instructional, persuasive, popular, reportage	600,000
Proficient L2 writing (Dutch as L1)	Journalistic texts	10,000
Proficient L2 writing (Afrikaans as L1)	Academic, instructional, popular, reportage	870,000

After extracting all relevant declarative complement clauses from these corpora, using verbs controlling finite declarative complement clauses as search strategy (based on the list in Quirk et al. (1985: 1180-1183), we coded the data for 7 language-internal and external factors, selected as operationalizations that allow us to disentangle the three proposed explanatory mechanisms for increased explicitness of lexicogrammatical encoding in translation. These factors include a.o. register (which can be seen as related to risk avoidance in respect of the degree of formality), source language structure (related to cross-linguistic influence) and distance between matrix verb of the main clause and the onset of the complement clause (related to cognitive effort). By measuring the relative effect of these factors, we are able to gain insights in the three explanations offered for increased explicitness, in different contact varieties.

The analysis method we adopt (and the major methodological contribution of this paper), is the so-called Multifactorial Prediction and Deviation Analysis (MuPDAR) method developed by Gries & Deshors (2015). This procedure represents an influential methodological advance in studying variation in language contact situations, where linguistic choices in a non-prototypical variety can be directly studied in relation to the central variety. More particularly, a first generalized linear mixed effects model is fitted on the data of the central variety only (British English), which reveals which factors significantly influence the choice between *that*- and zero-complementation, and to what extent. Subsequently, the outcome of this 'central' model is used to predict for each of the data points in the other varieties what British English language users would have done, and to what extent the choices in the contact varieties deviate from these.

Preliminary results show that the choices made in contact varieties are not altogether different from the central variety (yielding similarity scores of 76% and higher); however, some factor levels stimulate significantly different behavior in ways that reflect differential effects of contact in translation and L2 writing. Based on these findings, we re-evaluate the proposed increased explicitness of translated language through the frame of contact explanations, outlining the methodological advantages of multifactorial methods over frequency-based methods favored in earlier studies and demonstrating how CBTS benefits from interfacing both conceptually and methodologically with neighboring disciplines.

References

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

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

Bisiada, M. (2016). 'Lösen Sie Schachtelsätze möglichst auf': the Impact of Editorial Guidelines on Sentence Splitting in German Business Article Translations. *Applied Linguistics* 37(3), 354-76.

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

Gries, S. Th. & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9(1), 109-136.

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.

Hansen-Schirra, S. (2017). Between normalization and shining-through: Mixed methods for researching translation processes. Plenary lecture at *Translation and Interpreting in Transition,* Ghent University, 13-14 July 2017.

Kruger, H. (2012). A corpus-based study of the mediation effect in translated and edited language. *Target* 24(2), 355-88.

Kruger, H. (to appear). *That* again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*.

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.

Oakes, M. & Ji, M. (2012). *Quantitative Methods in Corpus-Based Translation Studies. A Practical Guide to Descriptive Translation Research*. Amsterdam & Philadelphia: Benjamins.

Olohan, M. & Baker, M. (2000). Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1(2), 141-158.

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

Serbina, T., Niemietz, P. & Neumann, S. (2015). Development of a keystroke logged translation corpus. In C. Fantinuoli & F. Zanettin (eds). *New Directions in Corpus-based Translation Studies*. Berlin: Language Sciences Press, 11-33.

Corpus-based analysis of style in students’ non-literary translations

Margherita Dore
University of Rome “La Sapienza”
margherita.dore@uniroma1.it

Style, defined as ‘the linguistic characteristics of a particular text’ (Leech & Short 2007: 11), is often conceived as the most important feature of literary texts. It is therefore not surprising that a wealth of scholarly research had been devoted to its analysis in Linguistics, Stylistics and Narratology (Leech & Short 2007; Semino 1997; Semino & Culpepper 2002, just to name a few). The relationship between style and translation has also been addressed by a number of scholars within Translation Studies, in both theoretical (Baker 2000; Boase-Beier 2006, 2011) and practical terms (Parks 2007). Indeed, both Stylistics and TS are concerned with a detailed linguistic analysis, specific textual choices, and their effects on readers. However, most analyses in both Stylistics and TS concentrate on literary work, although non-literary text-types also deserve thorough consideration in terms of style. Hence, since both fields focus on *what* is said but especially on *how* it is said, this study seeks to shed some light on translation, through stylistics (see Boase-Beier 2004 on ambiguity, or Marco 2004 on transitivity), and style, through translation (Boase-Beier 2011), by exploring specifically non-literary texts. The rationale behind this lies primarily in the fact that the concept of style seldom receives the attention it deserves, especially when the texts to be translated are other than fiction. Although it has been pointed out that translations of non-literary texts focus mainly on the message rather than the style (cf. Gutt 2000: 130 in Boase-Beier 2006: 27), the relevance of the latter cannot be underestimated. Genres such as museum and tourist guides display specific stylistic features that seek to convey the message in a very precise and appealing way. Such features may sometimes be underestimated, especially when the translator still has not developed the sufficient sensitivity to them. Hence, rather than concentrating on professional and/or official translations, the analysis will be based on a number of translation tasks completed by a group of third-year undergraduate students studying for a degree in English and Translation Studies at the University of Rome “La Sapienza” (B2 level and above). This can help to verify whether, and, if so, how, style is tackled by students who are approaching translation as part of their undergraduate training. In order to do this, the interdisciplinary quality of contemporary stylistics (its pragmatic, cognitive, sociological perspectives) will be tested in connection with translating strategies. This approach will be supported by a corpus-based analysis (Baker 1993, 1995) and learner corpus research (Granger 1993, 1994).

This pilot project includes two sample translations taken from different non-literary sources (i.e. museum and tourist guides), which have been translated from English into Italian by 34 students. Students were asked to deliver their translations within one week from the day they received the material via email. They were allowed to used CAT as well as traditional tools (e.g. monolingual and bilingual dictionaries) to complete their tasks. They were also asked to pay particular attention to the text type and terminology used. Students’ translations were then corrected against the official translation and marked accordingly.

Once completed, the translation database contained 68 translations which have been analysed via the corpus-based techniques offered by the Multilingual Student Translation (MUST) project (Granger & Lefer 2016). The use of this corpus-based approach allowed for the collection of standardised metadata. Moreover, thanks to this approach, recurrent patterns have been detected regarding language learners’ translation skills in relation to style. Lexical choices have been thoroughly investigated in connection with the text-type under scrutiny. The systematic analysis of the database has revealed that students only pay a limited amount of attention to the language choices that shape the style of a text. This lack of stylistic awareness may depend on the fact that undergraduate students are not sufficiently exposed to such textual features and the relevance they have within given genres. Consequently, it is suggested that much more emphasis should be placed on how non-literary source texts are first analysed and later translated, in such a way as to convey not only the message but also the style of the text, which is also directly linked to the genre of the text itself. A further step that is envisaged is the comparative analysis of the texts in this database with respect to the larger corpus available within the MUST project and platform. Also, didactic outputs may be developed in collaboration with all MUST project partners and contributors to enhance language teaching and translation training.

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: John Benjamins, 233-250.

Baker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research, *Target* 7(2), 223-243.

Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator, *Target* 12(2), 241-266.

Boase-Beier, J. (2004). Saying what Someone Else Meant: Style, Relevance and Translation. *International Journal of Applied Linguistics* 14(2), 276-287.

Boase-Beier, J. (2006). *Stylistic Approaches to Translation*. Manchester: St Jerome Publishing.

Boase-Beier, J. (2011). Stylistics and Translation. In Y. Gambier & Van L. Doorslaer (eds). *Handbook of Translation Studies* (Vol. 2). Amsterdam: John Benjamins, 153-156.

Catford, J. C. (1965). *A Linguistic Theory of Translation.* Oxford: Oxford University Press.

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan & N. Oostdijk (eds). *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam & Atlanta: Rodopi, 57-69.

Granger, S. (1994). The Learner Corpus: A Revolution in Applied Linguistics. *English Today* 39(10/3), 25-29.

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 88. Lund: Lund University Press, 37-51.

Granger, S. & Lefer, M.-A. (2016). The Multilingual Student Translation (MUST) project https://uclouvain.be/en/research-institutes/ilc/cecl/must.html (last accessed: 20.01.2018).

Gutt, E. A. (2000). *Translation and Relevance* (2nd edition). Manchester: St Jerome Publishing.

Leech, G. & Short, M. (2007/1985). *Style in Fiction*. London & New York: Longman.

Marco, J. (2004). Translating Style and Style of Translating: Henry James and Edgar Allan Poe in Catalan. *Language and Literature* 13(1), 73-90.

Parks, T. (2007). *Translating Style,* London: Routledge.

Semino, E. (1997). *Language and world creation in poems and other texts*. London: Longman.

Semino, E. & Culpepper, J. (eds). (2002). *Cognitive stylistics: language and cognition in text analysis*. Amsterdam: John Benjamins.

Connective placement in English and French: A cross-register corpus study

Maïté Dupont
Université catholique de Louvain
maite.dupont@uclouvain.be

In many languages, adverbial connectives such as *however, therefore* or *in addition* function as syntactically mobile elements, and may occur in a variety of positions in the sentence (see e.g. Biber et al. 1999: 890-2; Lenker 2014; Bonami et al. 2004; Grevisse & Goosse 2011: 1243; Altenberg 2006). Several contrastive studies have suggested that languages vary with respect to the positions that they tend to prefer within the range of options available to them. Altenberg (2006), for example, demonstrates that while English tends to use a majority of adverbial connectives in initial position, in Swedish the most frequent slot is the medial position. Based on a comparable corpus of newspaper editorials, Dupont (2015) uncovers marked differences in the preferred positions of English and French adverbial connectives of contrast, despite a similar set of possible positions in the two languages: English connectives are found to be used predominantly in initial position, whereas French connectives display a strong tendency to occur medially, within the verb phrase. Likewise, Balažic Bulc & Gorjanc (2015) show that Croatian connectives occur in initial position nearly twice as frequently as their Slovene equivalents in a comparable corpus of academic articles.

The few corpus-based studies comparing the placement patterns of adverbial connectives across languages, however, have made fairly general observations on the differences between the languages studied, without considering the possibility that those cross-linguistic contrasts may be influenced by the situation of communication. In line with the recent insistence, in corpus-based contrastive linguistics, on the importance of taking register variation into account when formulating differences between languages (Lefer & Vogeleer 2016; Neumann 2016), the objective of this paper is to investigate and compare the placement patterns of English and French adverbial connectives of contrast across two registers (viz. newspaper editorials and research articles) in order to assess the impact of register on cross-linguistic differences in connective placement. I will attempt to determine whether the differences between the two languages are stable across registers, or whether cross-linguistic differences in connective placement appear to be register-dependent.

The study is based on a comparable corpus made up of two subcomponents, each representing one language register: (i) the Mult-Ed corpus, which contains quality newspaper editorials compiled from newspapers including *The Guardian*, *The Observer*; *Le Monde*, *Le Figaro*, etc. (c. 2 million words per language); and (ii) the bilingual subpart of the Louvain Corpus of Research Articles (LOCRA), consisting of research articles from top-ranked journals across five disciplines in the Humanities (c. 2 million words per language). Based on a list of 32 English and 34 French connectives (obtained by pooling the inventories of connectives available in the literature), all the adverbial connectives of contrast were extracted automatically from the corpus with WordSmith Tools 6 (Scott 2012). The data was subsequently disambiguated manually in context, and only the connectives occurring at least 50 times (after disambiguation) per subcorpus were kept for further analysis (i.e. 8 connectives in English, and 7 in French). The manual annotation of position was based on the Systemic Functional notions of theme and rheme (Halliday & Matthiessen 2004) and distinguished between five positions, i.e. two within the theme, and three within the rheme.

The corpus results reveal a significant impact of register on connective placement in both English and French: in both languages, the editorials display a significantly higher frequency of connectives in rhematic positions – as in (1) and (2) – than the research articles, which are more strongly associated with the thematic positions – and more particularly the initial position, as in (3) and (4). However, while register is shown to generate variation in connective placement within each language system, it does not appear to have an influence on the cross-linguistic differences between English and French, which remain stable across the subcorpora: in both the editorials and the research articles, thematic connectives – as in (3) and (4) – and rhematic connectives used directly after the topical theme – as in (1) – are found to be more typical of English, whereas rhematic connectives used within the verb phrase – as in (2) – are more frequent in French. In other words, the cross-linguistic differences in connective placement do not seem to be register-dependent, and while both language and register play a significant role in the placement patterns of adverbial connectives of contrast, the preferences of each language appear to influence placement to a larger extent than the communicative situation in which the connectives are used.

(1) *Past premiers would give just the occasional speech and the even more occasional one-to-one interview […]. Mr Blair,* ***on the other hand****, has rewritten the handbook on prime ministerial accessibility and accountability* (Mult-Ed)
(2) *Le dispositif de Matthew Lipman a* ***cependant*** *subi […] des modifications plus ou moins importantes* (LOCRA)
(3) ***Néanmoins,*** *le Bonheur suprême maintint dans ces années-là son caractère franco népalais* (LOCRA)
(4) ***Nevertheless****, it is wrong for the Government to target food manufacturers* (Mult-Ed)

Such a conclusion is supported by the application of the multifactorial statistical method of classification and regression trees (CART) on the data: although the statistics reveal a significant impact of both language and register on connective placement, they highlight that language plays a more extensive role than register in explaining the differences emerging from the corpus. Interestingly, the statistical analysis of the data also makes lexis emerge as an influential factor for connective placement, with some connectives displaying idiosyncratic placement patterns within each language system.

In the final part of the study, a more qualitative approach to connective placement will be presented, focusing on the discourse effects that can be achieved by the choice of some specific positions for adverbial connectives of contrast. More particularly, it will be shown that connectives of contrast used in rhematic positions tend to fulfil a number of rhetorical functions – such as focusing attention on the theme, or acting as separators between given and new information – in addition to their basic linking function (see also e.g. Altenberg 2006; Lenker 2014). The analysis of the discourse effects produced by connective placement will be shown to provide additional insights into the cross-register variation uncovered in both the English and the French data.

References

Altenberg, B. (2006). The function of adverbial connectors in second initial position in English and Swedish. In K. Aijmer & A.-M. Simon-Vandenbergen (eds). *Pragmatic Markers in Contrast*. Oxford: Elsevier, 11-37.

Balažic Bulc, T. & Vojko G. (2015). The position of connectors in Slovene and Croatian student academic writing: a corpus-based approach. In S. Starc, C. Jones & A. Maiorani (eds). *Meaning Making in Text*. Basingstoke: Palgrave Macmillan UK, 51-71.

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

Bonami, O., Godard, D. & Kampers-Manhe, B. (2004). Adverb classification. In F. Corblin & H. de Swart (eds). *Handbook of French Semantics*. Stanford: CSLI, 143-184.

Dupont, M. (2015). Word order in English and French: the position of English and French adverbial connectors of contrast. *English Text Construction* 8(1), 88-124.

Grevisse, M. & Goosse, A. (2011). *Le bon usage*. Brussels: De Boeck-Duculot.

Lefer, M.-A. & Vogeleer, S. (2016). *Genre- and Register-related Discourse Features in Contrast*. Amsterdam: John Benjamins.

Lenker, U. (2014). Knitting and splitting information: Medial placement of linking adverbials in the history of English. In S. Pfenninger, O. Timofeeva, A.-C. Gardner, A. Honkapohja, M. Hundt & D. Schreier (eds). *Contact, Variation, and Change in the History of English*. Amsterdam: John Benjamins, 11-38.

Neumann, S. (2016). Cross-linguistic register studies: theoretical and methodological considerations. In M.-A. Lefer & S. Vogeleer (eds). *Genre- and Register-related Discourse Features in Contrast*. Amsterdam: John Benjamins, 35-57.

Scott, M. (2012). *WordSmith Tools*. Liverpool: Lexical Analysis Software.

What is right? Extra-clausal constituents in English and Swedish original and translated text

Anna Elgemark
University West
anna.elgemark@hv.com

The present study is an explorative corpus-based contrastive study of extra-clausal constituents (ECCs) at the right periphery in English and Swedish Fiction and Popular Science texts. ECCs are here defined as constituents outside the clause proper, but loosely associated with it in terms of pragmatic functionality (Dik 1997: 310). They provide additional information to clarify or modify either the whole clause, or part of it. When left out of the clause, the remaining clause structure is still complete and grammatical. Typically, they are marked off by punctuation in writing and intonation in speech. In this position, we find different types of constituents in a range of definitions: Afterthought, Appendage, Apposition, Right dislocation, Tag and Tail to name a few.[1]

The present study builds on Elgemark (2017), in which ECCs at the right periphery were shown to have different characteristics in English and Swedish.[2] In Swedish, they are to a large extent related to an NP/Subject Theme; they could be seen as Substitute Themes (Matthiessen 1995: 563), enabling the presentation of a participant both as Theme (light, ensuring topic continuity) and as focus at the end of the clause, as is illustrated in (1) and (2):[3]

(1) but it had been one of the things that attracted Marjorie when they bought the house two years ago – **the bathroom with its kidney-shaped hand basin and gold plated taps and sunken bath and streamlined loo and bidet.** (DLO1: 77)

(2) *Vi* tog tåget, ***pappa och Siiri och jag.*** (AP1: 222)
'We took the train, **dad and Siiri and me.'**

However, in English, ECCs in post-clausal position are to a greater extent oriented to the Process, frequently realised as Non-Finite *–ing* clauses used for syntactic compression and/or to emphasise the simultaneity of the actions presented in the *–ing* clause and the preceding main clause, as is illustrated in (3).

(3) We would stand side by side, **looking at a large red mouth stretching itself around a chocolate bar**, (MA1: 195)

Elgemark (2017) also showed that there was low translation correspondence between ECCs[4] in the two languages. This indicates that there are contrastive differences regarding the types of constituents found at the right periphery in the two languages. Therefore, the present study sets out to further examine ECCs in the two languages by focusing primarily on translated English and Swedish. Thus, the aim is to explore the characteristics of ECCs at the right periphery in English and Swedish, in original as well as translated texts. In accordance with this aim, the following research questions will be addressed:

- ✓ What are the typical formal and functional properties of ECCs at the right periphery in original and translated English and Swedish texts?
- ✓ To what extent is there correspondence between ECCs in original and translated English and Swedish?
- ✓ How could the lack of correspondence be explained?

The analysed ECCs have been extracted from the Fiction and Non-Fiction part of the English-Swedish Parallel Corpus (ESPC), which consists of comparable original texts in English and Swedish as well as their translations

[1] Teleman et al. (1999 (4): 438-458) uses the term 'Annex' for dislocated constituents, sentence adverbials, appositions placed in clause-final or clause-initial positions.
[2] In Elgemark (2017), these constituents are referred to as Tail.
[3] In examples (1) and (2), the referent of the extra-clausal constituent has been underlined.
[4] Tail in Elgemark (2017).

into the other language. Thus, it has the advantage of being both a comparable and a translation corpus. The size of the corpus is 2.8 million words and its structure is shown in Figure1:

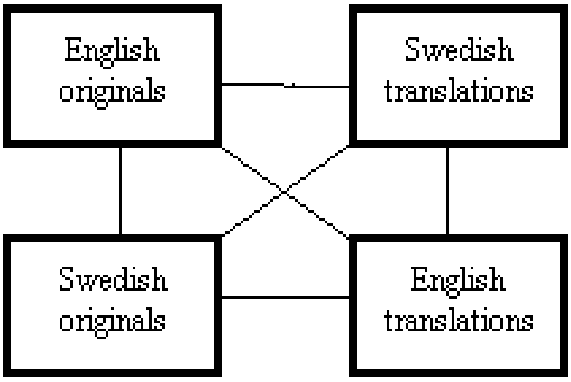


Figure 1. Structure of the English-Swedish Parallel Corpus (Altenberg et al. 2001)

Many of the constituents found in the right periphery are particularly frequent in spoken language. Often they are used as repairs to clarify a referent in the preceding context that the speaker assumes to be unclear. In view of this, ECCs in Fiction as well as Popular Science texts have been analysed, as we might find different types of ECCs with different functions in the two text types.

The present study uses the corpora both as a comparable corpus and as a translation corpus. A comparison of original and translated texts in the same language could reveal systematic differences, referred to as translation universals by Baker (1992, 1996).[5] The universals include *simplification*, *explicitation, normalisation or conservatism*, and finally, *levelling out* (Baker 1996: 176f). Furthermore, Teich (2003) highlights two ways in which translations are different from source texts in the same language. First, they bear resemblances of the original text, SL shining through, and second, they try to be more typical of the target language than original texts in the same language. These two contradictory processes work at the same time affecting different parts of the language (2003: 219). Similarly, Gellerstam (1985: 88) defines the 'systematic influence on target language (TL) from source language (SL), or at least generalizations of some kind based on such influence', as Translationese.

The extent to which the above-mentioned processes could actually be seen as universals has been questioned by some scholars (see e.g. Tirkkonen-Condit 2002; House 2008). House (2008: 11) emphasises the fact that translations always are language-specific. Thus, it is problematic to claim that explicitation taking place in translations between e.g. English and Swedish is an indication of a universal phenomenon. Rather, it seems to be a feature of this specific translation pair. Similarly, Becher (2011) emphasises that it is important to try to 'trace as many occurrences of explicitation as possible back to lexicogrammatical and pragmatic differences between the source and target language' (2011: 14).

In view of this, the present study attempts to integrate contrastive linguistics and translation theory in the analysis of ECCs at the right periphery in English and Swedish original and translated texts. This could gain further insights into the characteristics of the two languages, as well as the effects of the translation process.

[5] For a discussion of translation universals, see e.g. Blum-Kulka (1986), House (2008), Laviosa (2002), Malmkjaer (2005), Mauranen & Kujamäki (2004) and Toury (1995).

References

Altenberg, B., Aijmer, K. & Svensson, M. (2001). *The English-Swedish Parallel Corpus: Manual of Enlarged Versions.* Department of English, University of Lund and Department of English, University of Gothenburg. Available at http://www.sol.lu.se/engelska/corpus/corpus/espc.html. Last accessed January 2018.

Baker, M. (1992). *In Other Words. A coursebook on translation*. London & New York: Routledge.

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.

Becher, V. (2011). *Explicitation and Implicitation in Translation. A corpus-based study of English-German and German-English translations of business texts*. PhD thesis, Universität Hamburg.

Carter, R. & McCarthy, M. (2006). *Cambridge Grammar of English. A Comprehensive Guide*. Cambridge: Cambridge University Press.

Dik, S. C. (1997a). *The Theory of Functional Grammar. Part 1: The structure of the clause.* Berlin & New York: Mouton de Gruyter.

Elgemark, A. (2017). *To the Very End. A contrastive study of N-Rhemes in English and Swedish translations*. Department of Languages and Literatures, University of Gothenburg.

Gellerstam, M. (1985). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (eds). *Translation studies in Scandinavia. Proceedings from The Scandinavian symposium on translation theory (SSOTT) II,* Lund 14-15 June. Lund: Gleerup, 88-95.

House, J. (2008). Beyond Intervention: Universals in Translation? *trans-kom* 1(1), 6-19.

Huddleston, R. D. & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press.

Matthiessen, C. M. I. M. (1995). *Lexico-Grammatical Cartography*: *English systems*. Tokyo: International Language Sciences Publishers.

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

Tirkkonen-Condit, S. (2002). Translationese. A myth or an empirical fact? A study into the linguistic identifiability of translated language. *Target* 14(2), 207-220.

Translation, genres, and source languages: Searching for phraseological regularities from the bottom up

Adriano Ferraresi, Ilmari Ivaska, Silvia Bernardini
University of Bologna
adriano.ferraresi@unibo.it, ilmari.ivaska@unibo.it, silvia.bernardini@unibo.it

Introduction

This contribution focuses on phraseology in translated English, aiming to identify differences in the use of collocations that set apart translated from non-translated texts. The study sets itself against the combined background of the neo-Firthian frequency-based view of collocations as combinations of words that occur together more often than predicted by chance (e.g. Jones & Sinclair 1974/1996; Evert 2005), and of the substantial work carried out in corpus-based translation studies to identify typical features of translated language (e.g. Baker 2007).

Previous work in this area has uncovered several ways in which translations seem to differ from same-language originals (Laviosa 2002). It has also pointed out that other variables, such as genre- and source language-related effects, should also be taken into account, since they may play a confounding role when trying to tease apart more volatile differences along the translated/non-translated dimension (Koppel & Ordan 2011). The focus on collocations is in turn motivated by the fact that, being restricted by use, phraseological regularities are among those "properties of discourses (…) which are below the threshold, or outside the realm of, borrowing of lexical or structural patterns across languages" (Steiner 2008: 321). The relative presence/absence of different types of word combinations in translated texts (e.g. in a frequency-based approach, unusual vs. strongly associated word combinations) may provide insights about typical features of translation, such as interference or normalization/sanitization (Kenny 2001).

Aims and method

The basic question we address in this study is whether translated English texts differ from comparable texts originally written in English in terms of the phraseological patterns they use. In an attempt to disentangle the effects of unrelated variables, our purpose-built corpus contains translations from two source languages and comparable original English texts in three genres: (transcriptions of) EU parliamentary speeches, news texts, and tourist guides, for a total of 900 texts in 9 components (426,112 tokens overall). The two source languages, Italian and Finnish, are typologically distant from each other and from English, making it more straightforward to detect linguistic interference effects, if any.

Differently from most previous work on collocations in monolingual comparable corpora (e.g. Durrant & Schmitt 2009), we do not preselect patterns based on intuition or previous research (but see e.g. Granger & Bestgen 2015); rather we adopt a corpus-driven method, and apply it to syntactically parsed data. Our corpus is annotated using the UDPipe parser (Straka & Straková 2017). Frequencies of all continuous or discontinuous sequences of two Parts Of Speech (POS) are extracted relying on the dependency structure assigned by the parser, and normalized over 1,000 tokens for each text. We then select POS sequences whose elements are lexical, and use random forests (see e.g. Tagliamonte & Baayen 2012) to rank them in terms of how well they distinguish between translated and non-translated texts, and to compare the relative importance of genre and source language status as alternative predictors.

In a second stage, we select the POS sequence that best predicts translated status, extract the corresponding lemma pairs (almost 11,000), and skim through the list to discard malformed ones resulting from parsing/lemmatization errors (less than 2% of the total number of pairs). To assess the phraseological nature of the pairs, we classify them based on three lexical association measures, i.e. a) raw frequency, b) Mutual Information and c) t-score, deriving frequency data from ukWaC (Baroni et al. 2009) to ensure greater reliability of the scores (see Durrant & Schmitt 2009). Pairs with a frequency of 0 in ukWaC are classified as *unattested*, while those scoring high on both MI and t-score (i.e. whose values are above the median values for attested pairs; MI > 0.9, t-score > 2.3) are classified as *highly salient* collocations. The number of combinations belonging to the two categories is then calculated for each text and expressed as a percentage.

Preliminary results

At the syntactic level, the five best predictors of the translated/non-translated status of a text are NOUN(node)–VERB(head), NOUN(node)–NOUN(head), NOUN(head)–NOUN(node), VERB(head)–VERB(node), and VERB(head)–NOUN(node). When these results are set against those for the genre and source language comparisons, frequencies are found to consistently predict either the genre or the source language better than the translated status (differences are greater), confirming the confounding nature of these factors for investigations into the typical features of translated language.

The lexical POS pair that best predicts translated/non-translated status is NOUN(node)–VERB(head), i.e. a structure in which a noun precedes and syntactically depends on a verb. An example is DELAY–MEASURE in "The delay in translating those papers into eleven different languages can be measured in terms of extra deaths". This structure is twice as good a predictor of translated-ness as the second-best predictor, NOUN(node)–NOUN(head). This POS pair is more frequent in translated than non-translated language, though the difference is particularly marked for Finnish as a source language in the news genre (Figure 1).

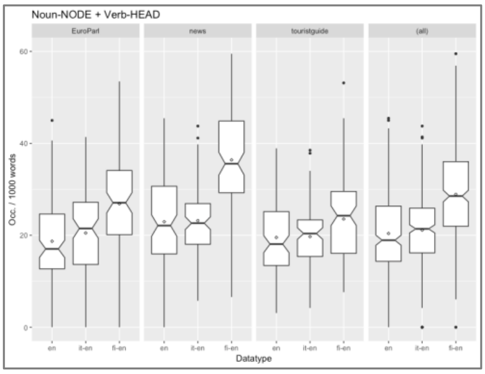


Figure 1. NOUN(node)–VERB(head) tokens

When factoring in the phraseological status of the NOUN(node)–VERB(head) pairs (Figure 2), no significant differences emerge between the subcorpora of parliamentary proceedings: the frequent use of the structure in this genre does not appear to be related to the use of specific types of word combinations (i.e., unattested pairs or highly salient collocations). In the news and tourist guide genres, we do observe significant differences, which however only concern translations from Finnish with respect to original English. While English translated from Italian and original English make use of similar numbers of both types of pairs, translations from Finnish are richer in highly salient collocations and poorer in unattested pairs, pointing at greater phraseological conventionality.

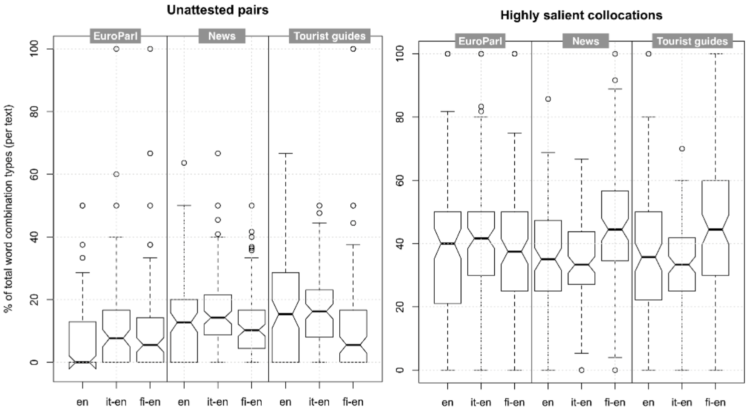


Figure 2. Unattested pairs and highly salient collocations in the NOUN(node)–VERB(head) pattern

Conclusion

In this contribution we have described a corpus-driven method to single out phraseologically productive syntactic patterns that distinguish translated and non-translated language. This method has been applied to a genre- and source language-controlled dataset. Focusing on the most distinctive pattern, we have compared the relative frequency of highly salient and unattested word combinations across the different subcorpora. The ensuing phraseological analysis has revealed that English translated from Finnish is characterized by greater phraseological conventionality of this structural pattern compared to non-translated English in two out of three genres. No such difference was observed in translations from Italian, nor in the genre of EU parliamentary proceedings.

The paper will discuss possible explanations and implications of these findings, and reflect on the potential of our method to detect salient features of translated language at different levels of analysis, while keeping track of relevant contextual variables.

References

Baker, M. (2007). Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics* 21(1), 11-21.

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.

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

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations.* PhD thesis, University of Stuttgart.

Granger, S. & Bestgen, Y. (2015). Using collgrams to assess L2 phraseological development: A replication study. In P. de Haan, R. de Vries & S. van Vuuren (eds). *Language, Learners and Levels: Progression and Variation*. Louvain-la-Neuve: Presses universitaires de Louvain, 385-408.

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

Koppel, M. & Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the ACL*. Portland, Oregon: Association for Computational Linguistics, 1318-1326.

Laviosa, S. (2002). *Corpus-based Translation Studies*. *Theory, findings, applications.* New York: Rodopi.

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: Benjamins, 317-345.

Straka, M. & Straková, J. (2017). Tokenizing, POS Tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of CoNLL 2017 shared task: Multilingual parsing*. Vancouver, Canada: Association for Computational Linguistics, 88-99.

Tagliamonte, A. S. & Baayen, R. (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24, 135-178.

What makes human translation different?

Ana Frankenberg-Garcia
University of Surrey
a.frankenberg-garcia@surrey.ac.uk

When Warren Weaver proposed the ideal of machine translation (MT) in the post-war era (Weaver 1949), it was hard to conceive that anything near human translation (HT) would one day become a reality. However, after a long period of frustrated attempts to overcome the limitations of rule-based MT, the proliferation of digital texts in the nineties and the resulting availability of more and more human translations that could be used as training data paved the way for the development of phrase-based, statistical MT (Koehn et al. 2003). This represented a huge leap in the quality of MT texts. Then 2016 marked the beginning of a new paradigm in the quest for perfecting MT, with the neural approach taking lead (Luong et al. 2016). For language pairs where sufficient training data exists, the initial results of neural MT we are seeing today are truly remarkable.

While the development of MT, whether phrase-based or neural, still depends on large amounts of training data from HT, when comparing MT and HT attention has been paid mainly to assessing the quality of MT output, where the closer MT is to HT, the better it is. However, given that MT operates predominantly at the level of the phrase or sentence, less attention has been paid to discourse (Hardmeier 2012). While there is some research on MT and certain features of discourse such as lexical cohesion, terminological consistency, use of connectives and pronoun prediction (e.g. Carpuat & Simard 2012; Cartoni et al. 2011; Guillou 2013; Hardmeier et al. 2015; Meyer & Weber 2013; Russo et al. 2011; Voigt & Jurafsky 2012; Weber et al. 2017), the focus of this research has been on the development of highly-specific algorithms to improve MT output. In this paper we address this question from the perspective of Translation Studies instead. Our primary aim is to come to a better understanding of discourse features that differentiate human translators.

Unlike the majority of MT research, where comparison with HT typically disregards the conditions under which HT was achieved, the present study examines the work of human translators carrying out authentic translation tasks. Moreover, the study controlled for (1) the directionality of translation, (2) a balanced representation of different authors and different translators, (3) the possible influence of MT upon HT, and (4) the constraining effect of computer-assisted translation (CAT) tools on discourse.

The starting point to this investigation was the creation of a source-text corpus of Portuguese aligned with English HT. The sampling was opportunistic, making use of existing data from the COMPARA parallel corpus of Portuguese and English fiction (Frankenberg-Garcia & Santos 2003). To ensure that the sub-corpus selected for this study was balanced, a sample of full-sentence concordances adding up to between four and five thousand source-texts words from the work of fifteen published Portuguese-speaking authors translated by fifteen different professional English translators was downloaded from the online interface to COMPARA (for copyright reasons, the system randomly restricts the number of concordances that can be retrieved for each bitext).

To obtain the MT output, the Portuguese source-text segments of the corpus were machine translated into English using the Google Translator Toolkit. Google uses advanced neural MT technology for the Portuguese-English language pair (Turovsky 2016), and its output is generally recognized as very good, especially in the morphologically-rich-to-poor Portuguese into English direction. With the resulting English MT output, it was possible to arrive at three perfectly aligned corpora:

1. Portuguese ST corpus of circa 72K words (from COMPARA)
2. English HT corpus of circa 82K words (from COMPARA)
3. English MT corpus of circa 77K words (from Google Translate, 2017)

As fiction is not a genre that has so far been aided by MT and MT engines are not customarily trained on the translation of fiction (Toral & Way 2018), it can be assumed that the HT and MT corpora were not cross-contaminated. Likewise, as the translation of fiction is not normally carried out in a CAT environment, and as the English translations in COMPARA are authentic published translations dating back to the eighties and nineties – a time before the use of CAT tools became widespread – it can also be assumed that the HT in the corpus were not constrained by such tools.

The three corpora were compiled as a parallel corpus on Sketch Engine (Kilgarriff et al. 2004), allowing one to carry out three-dimensional parallel queries – ST-HT-MT – and navigate each corpus separately in order to explore how HT and MT differed in terms of discourse. While a few of the discourse features examined were informed by previous work in Translation Studies – for example, sentence splitting and joining, the use of conjunctions, and position of adverbs (Frankenberg-Garcia 2014, 2019) – others were corpus-driven, starting from a keyword analysis comparing discrepant frequencies in the HT and MT corpora, and then observing parallel concordances to examine whether those discrepancies had implications at the level of discourse.

Preliminary findings indicate that human translators tend to write longer sentences, use more varied vocabulary, use more coordinating conjunctions and possessive pronouns. A first-hand account of the main discourse differences between HT and MT detected in this investigation will be presented at UCCTS.

References

Carpuat, M. & Simard, M. (2012). The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal (Canada), 442-449.

COMPARA corpus (1999-2008). Online: http://www.linguateca.pt/COMPARA/index.php [18/12/2017].

Frankenberg-Garcia (2014). Understanding Portuguese translations with the help of corpora. In T. Sardinha & T. Ferreira (eds). *Working with Portuguese Corpora*. London: Bloomsbury, 161-176.

Frankenberg-Garcia, A. (2019). A corpus study of splitting and joining sentences in translation. *Corpora*, 14(1).

Frankenberg-Garcia, A. & Santos, D. (2003). Introducing COMPARA: the Portuguese-English Parallel Corpus. In F. Zanettin, S. Bernardini & D. Stewart (eds). *Corpora in Translator Education*. Manchester: St. Jerome, 71-87.

Google Translator Toolkit (n.d.). Online https://translate.google.com/toolkit [18/12/2017].

Guillou, L. (2013). Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, Soa (Bulgaria), 10-18.

Hardmeier, C. (2012). Discourse in statistical machine translation: a survey and a case study. *Discours*, 11. Online: http://discours.revues.org/8726 [18/21/2017].

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J. Versley, Y. & Cettolo, M. (2015). Pronoun-focused MT and crosslingual pronoun prediction: findings of the 2015 DicoMT shared task on pronoun translation. In *Proceedings of DiscoMT 2015*, Lisbon (Portugal), 1-16.

Kilgarriff, A., Rychly, P., Smrz, P. &Tugwell, D. (2004). The Sketch Engine. *Proceedings of Euralex*. Lorient, France, 105-116.

Koehn, P., Och, F. & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton (Canada), 48-54.

Luong, T., Cho, K. & Manning, C. (2016). Neural Machine Translation. *ACL 2016 Tutorial*. https://sites.google.com/site/acl16nmt/ [18/12/2017].

Toral, A. & Way, A. (2018). What level of quality can neural machine translation attain on literary text? https://arxiv.org/abs/1801.04962 [11/03/2018].

Turovsky, B. (2016). Found in translation: More accurate, fluent sentences in Google Translate. https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/ [18/12/2017].

Voigt, R. & Jurafsky, D. (2012). Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, Montreal (Canada), 18-25.

Weaver, W. (1949) *The Weaver Memorandum*. Online: http://www.mt-archive.info/Weaver-1949.pdf [18/12/2017].

Webber, B., Popescu-Belis, A. & Tiedemann, J. (2017). *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen (Denmark), Association for computational Linguistics. http://www.aclweb.org/anthology/W/W17/W17-4800.pdf [18/12/2017].

The arrows move the selection – *Durch die Pfeile bewegen Sie die Auswahl*: A corpus-based study of non-agentive constructions in translation from English to German across three registers

Jonas Freiwald
RWTH Aachen University
jonas.freiwald@ifaar.rwth-aachen.de

Despite their strong genetic relation, the English and German languages show a variety of contrastive differences which pose problems during the translation process. One such contrastive difference is the use of non-agentive constructions. Non-agentive constructions are composed of an inanimate Subject, which means that it is at a low rank on the animacy scale (Zaenen et al. 2004) and an agentive Verb, which is a process that requires an entity that is capable of acting. Corpus studies by Biber et al. (2007) have shown that 30% of all agentive verbs in English are paired with inanimate Subjects, making such non-agentive constructions quite frequent in English. As English has lost most of its inflectional morphology, word order has also become stricter, with most grammatical functions being tied to specific positions in the clause. In declaratives, the Subject typically occupies the first position in the clause, which is also the default position for given information. As a consequence, the semantic mapping onto the Subject is very broad in English (König & Gast 2009), so that word order can align with the desired information structure.

The rich inflectional system of German expresses grammatical relations and thus allows a freer word order. With the exception of the finite Verb, which typically occupies the second position in the clause, all other clause elements can be moved around freely in the clause; thus, information structure does not have to correspond to any particular order of clause constituents. For this reason, there is a more direct relationship between grammatical functions and semantic mapping, which restricts the use of non-agentive constructions in German (Hawkins 1986).

Since non-agentive constructions are not strictly ungrammatical in German, translators, when faced with such a construction in the English original, have to make a choice between keeping the original sentence structure intact or changing it to make the translation more authentic in the target language. Based on these considerations, two hypotheses can be postulated:

1. Non-agentive constructions in the English originals will undergo translation changes more often than agentive constructions.
2. The translation of non-agentive constructions will display a wider variety of translation strategies.

If the original is changed, the translator has to decide to either deviate from the information structure or the grammatical structure of the original, or change the lexical or grammatical form of the Verb. Drawing on previous studies, this paper identifies four translation strategies that are likely to occur:

<u>Strategy 1</u>: Information structure stays intact, EO Subject changes grammatical function.

EO: Good, albeit patchy, evidence suggests that transposons contribute to the evolution and genomic regulation of higher organisms […].

GT: *Guten, wenn auch bisher lückenhaften Indizien nach tragen Transposons zur Evolution und genomischen Regulation höherer Organismen bei [...].*
(According to good, albeit patchy, evidence, transposons contribute to the evolution and genomic regulation of higher organisms […].)

<u>Strategy 2</u>: Grammatical structure stays intact, EO Subject is changed to an animate agent.

EO: Jaffe's vibrating probe confirmed that this was caused by electrical currents similar to those in Fucus.

GT: *Jaffe konnte hierfür mit Hilfe seiner Vibrationssonde ähnliche elektrische Ströme wie im Fucus nachweisen.*
(Jaffe was able to find currents similar to those in Fucus with the help of his vibrating probe.)

Strategy 3: Grammatical structure stays intact, EO Verb is changed to a non-agentive verb

EO: Fluoride salts dissolved in water show only weak chemical activity […].
GT: *In Wasser gelöste Fluoridsalze sind kaum chemisch aktiv [...].*
(Fluoride salts dissolved in water are hardly chemically active.)

Strategy 4: Information structure and grammatical structure stay intact, EO Verb is passivized

EO: This page doesn't show if you chose not to use the KDE Address Book.
GT: *Diese Seite wird nicht angezeigt, wenn Sie das KDE-Adressbuch nicht zur Speicherung von Kontakten verwenden.*
(This page is not shown if you don't use the KDE Address Book for saving contacts.)

All corpus results will be based on the CroCo Corpus, a bi-directional translation corpus of English and German. The CroCo corpus includes eight different registers and offers various annotation levels such as part-of-speech tagging and grammatical function analysis (Hansen-Schirra et al. 2012). The effects of non-agentive constructions on translations into German have already been analyzed in the register of popular scientific texts (Serbina 2015; Freiwald 2015). Drawing on these findings, this paper will extend the corpus analyses to three other registers, namely fictional texts, instruction manuals and tourism leaflets, to examine the relationship between contrastive differences and register characteristics. This selection of registers not only represents a variety of specialized and non-specialized language but also promises to be particularly relevant in regard to non-agentive constructions.

The corpus analysis will consist of annotations of Subject animacy and Verb agency for all three registers. To my knowledge, there is no automatic tool that reliably analyzes agentivity, which is why all annotation will be carried out manually using the UAM Corpus Tool (O'Donnell 2008). While the study will focus on non-agentive constructions in translations from English to German, it will also include a sample analysis of German originals in the same three registers to serve as a baseline for German texts. The results of these analyses will be tested statistically with the help of significance testing and logistic regression.

This paper will not only investigate the influence of non-agentive constructions on the translation product but also provide a detailed discussion of the most common translation strategies in the different registers. These results will help us to better understand both contrastive and register differences and deepen our general understanding of the translation behavior in translations between English and German.

References

Ahearn, L. (2001). Language and Agency. *Annual Review of Anthropology* 30, 109-137.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (2007). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Freiwald, J. (2015). *You say Theme, I say Thema*: A Corpus-Based Approach to Theme in English and German from an SFL Perspective. Staatsarbeit: RWTH Aachen.

Hansen-Schirra, S., Neumann, S. & Steiner, E. (2012). *Cross-Linguistic Corpora for the Study of Translations*. Berlin: De Gruyter.

Hawkins, J. (1986). *A Comparative Typology of English and German: Unifying the Contrasts*. London & Sydney: Croom Helm.

König, E. & Gast, V. (2009). *Understanding English-German Contrasts*. Berlin: Erich Schmidt.

O'Donnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. In C. M. Bretones Callejas, J. F. Fernández Sánchez, J. R. Ibáñez Ibáñez, M. E. García Sánchez, M. E. Cortés de los Ríos, S. Salaberri Ramiro, M. S. Cruz Martínez, N. Perdú Honeyman & B. Cantizano Márquez (eds). *Applied Linguistics Now: Understanding Language and Mind [La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente]*. Almería: Universidad de Almería, 1433-1447.

Serbina, T. (2015). *A Construction Grammar approach to the analysis of translation shifts: A corpus-based study*. PhD thesis. RWTH Aachen University, Aachen. Retrieved from https://publications.rwth-aachen.de/record/538325. January 2018.

Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M. C. & Wasow, T. (2004). Animacy Encoding in English: Why and How. In *Proceedings of the 2004 ACL workshop on discourse annotation*, Barcelona, Spain, 118-125. Retrieved from http://homepages.inf.ed.ac.uk/jeanc/acl-da-wkshop.pdf. January 2018.

MUST: A collaborative corpus collection initiative for translation teaching and research

Sylviane Granger, Marie-Aude Lefer
Université catholique de Louvain
sylviane.granger@uclouvain.be, marie-aude.lefer@uclouvain.be

Multilingual Student Translation (MUST) is a collaborative corpus collection initiative that brings together translation and foreign language teachers and researchers around two main objectives: to collect and share translations produced by students and to process them using a standardized set of tools and guidelines with a view to optimizing translation teaching and advancing empirical research.

Learner translation corpora, i.e. corpora containing translations produced by learners, are situated at the interface between learner corpus research (Granger et al. 2015) and translation studies (Malmkjaer 2017). As pointed out by Bowker & Benison (2003: 103), there is every reason to expect that the insights gained from collecting and analysing foreign language learner data could be of equal benefit to the field of translation studies: "Student translators can be considered as a highly specialized type of language learner/user. Although their specific needs differ from those of general language learners, a similar approach to collecting and studying the output of student translators would be highly valuable for both pedagogical and research applications".

The idea of compiling corpora of student translations is not new. Among the forerunners were Uzar & Walinski (2001), Bowker & Bennison (2002), Florén (2006) and Kübler (2007). More recent projects include Štěpánková (2014), Kutuzov & Kunilovskaya (2014) and Wurm (2016). A survey of learner translation corpora to date reveals a certain number of strengths, including the creation of error annotation systems tailor-made for translation, but also some weaknesses, in particular a limited number of language pairs and a tendency to restrict the corpus to one translation direction. In addition, the diversity of error annotation systems used makes it difficult to compare the results across corpora. More generally, most of the corpora are not available and, as observed by Espunya (2014: 35), "the field is clearly in its infancy, judging by the scarcity of publications reporting results or even research programmes". The MUST initiative was launched in 2016 with a view to filling these gaps. The collected data will be truly multilingual and represent a large number of text types, genres, registers and topics. Both L2>L1 and L1>L2 translations will be included. The project currently includes 30 research teams from 14 countries and covers a large number of languages (Chinese, French, Dutch, English, Galician, German, Greek, Italian, Lithuanian, Macedonian, Norwegian, Polish, (Brazilian) Portuguese, Slovene, Spanish). To ensure easy access for all partners in the project, all the data are collected and searchable on a web-based interface, Hypal4MUST, an adapted version of the Hypal interface designed by Obrusnik (2014) for the processing of parallel texts.

The presentation will give a general overview of the project and focus more particularly on two of its key features: tailor-made metadata and annotation, both of which are standardized and will be used for all the translations in the database in order to ensure full comparability of data and reliable interpretation of results.

Three layers of **metadata** are collected: (1) source-text-related metadata (e.g. genre & sub-genre, domain, mode, target audience, sampling); (2) translation-task-related metadata (e.g. type of task, marking, tools and resources, feedback and revision, use of a reference translation memory or terminology database); (3) translator-related metadata (e.g. language background, prior and current study background, self-rated proficiency in source and target language, translation experience). Once a text has been uploaded with all its metadata, it can be reused by other members of the consortium, thereby greatly facilitating their teaching and/or research activities, and effectively turning the interface into a truly collaborative platform.

The **Translation-Oriented Annotation System** (TAS) created within the framework of the project presents two distinctive characteristics: first, it offers the possibility of highlighting both erroneous and correct use; second, it offers the option of marking translation procedures (such as transposition, simplification or explicitation), thereby catering for theoretically oriented research. In view of these two features, it was decided to refer to the annotation system as "translation-oriented annotation" rather than "error annotation". The system contains 60 tags in total and is made up of the following three parts:

- Source text-target text transfer (TR): discrepancies between the source text and the target text and/or between the target text and the translation brief (e.g. distortion of the message: *policies that* can lead *to sustainable wellbeing* > *politiques qui* mènent *à un bien-être durable* 'policies that lead to sustainable wellbeing')
- Language of target text (LA): features of the target text that are erroneous and/or inappropriate independently of the source text (e.g. *the people which* instead of *the people who).*
- Translation procedures (TP): procedures used to solve translation problems, which can be observed when comparing the translation with its source text (e.g. implicitation, generalization, borrowing).

As it is hierarchical, the system allows analysts to tag at the level of granularity that best fits their teaching or research aims. They may opt for broad categories (in LA: grammar, lexis and terminology, cohesion, mechanics, style and situational context) or for more detailed annotation (e.g. distinguishing between spelling, punctuation and units/dates/numbers in the Mechanics category). TAS also contains two meta-tags that can be added to any portion of the text: a plus sign tag (Plus) to mark positive features, i.e. particularly good translation choices, and an SLI tag to highlight suspected source language intrusion features.

The presentation will also feature screenshots illustrating the workflow of the Hypal4MUST student and teacher web interfaces used to compile and search the database.

References

Bowker, L. & Bennison, P. (2002). Translation Tracking System: A tool for managing translation archives. *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, 29-31 May 2002, 503-507.

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation* 48(1), 33-43.

Florén, C. (2006). ENTRAD, an English Spanish parallel corpus created for the teaching of translation. Paper presented at the *Seventh Teaching and Language Corpora Conference* (TALC 2006).

Granger, S., Gilquin, G. & Meunier, F. (eds). (2015). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

Kübler, N. (2008). A comparable Learner Translator Corpus: Creation and use. *LREC 2008 Workshop on Comparable Corpora*, 73-78.

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

Malmkjaer, K. (2017). *The Routledge Handbook of Translation Studies and Linguistics.* Abingdon & New York: Routledge.

Obrusnik, A. (2014). Hypal: A User-Friendly Tool for Automatic Parallel Text Alignment and Error Tagging. *Eleventh International Conference Teaching and Language Corpora*, Lancaster, 20-23 July 2014, 67-69.

Štěpánková, K. (2014). Learner Translation Corpus: CELTraC (Czech-English Learner Translation Corpus). Bachelor's Diploma Thesis, Masaryk University.

Uzar, R. & Waliński, J. (2001). Analysing the fluency of translators. International Journal of *Corpus Linguistics* 6, 155–166.

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

Emotionality in Swedish and Polish: A parallel-corpus-based study

Ewa Gruszczyńska, Agnieszka Leńko-Szymańska, Anna Sworowska
University of Warsaw
e.gruszczyńska@uw.edu.pl, a.lenko@uw.edu.pl, asworowska@uw.edu.pl

Emotions have long been the centre of both theoretical and empirical inquiry in many academic disciplines including psychology, sociology, cognitive science and linguistics (e.g. Coutler 1986; Ortony et al. 1988; Ekman 1992; Wierzbicka 1992, 1994; Plutchik 1994). They have become a testing ground in the debate on the universality of human experience. One camp in this debate, the relativists, maintain that languages and cultures differ in the way they interpret, evaluate and express emotions (Wierzbicka 1988).

This paper focuses on exploring and comparing the linguistic representation of one emotion – ~'fear' – in Swedish and Polish. Its aim is to investigate whether there is a difference in the strength of emotional loading between Swedish original items referring to fear and their Polish equivalents in translated texts. While the linguistic representation of emotions in different languages and cultures has been widely discussed in literature (e.g. Athanasiadou & Tabakowska 1998; Harkins & Wierzbicka 2001; Hurtado de Mendoza 2008; Lewandowska-Tomaszczyk & Wilson 2013; Lüdtke 2015), this paper reports on one of the first studies addressing this issue in the context of translation, in particular between Swedish and Polish (see Shields & Clarke 2011 and Gruszczyńska 2001 for counterexamples).

Geert Hofstede (2001) demonstrated that the Swedish and Polish cultures differ significantly from each other on three dimensions: POWER DISTANCE, UNCERTAINTY AVOIDANCE and MASCULINITY. It can be expected that in result of these differences the two cultures differ along yet another dimension – that of EMOTIONALITY. Indeed, Swedes are stereotypically perceived as restrained in showing emotionality (cf. Daun 1989), and Poles as a nation with inclinations to frequent and unrestrained expression of emotions (Wierzbicka 1990). Thus, it has been hypothesised that the linguistic items which represent the basic emotion ~'fear' may be translated from Swedish into Polish with equivalents which are stronger in their emotional loading and vice versa from Polish into Swedish in a more subdued way, i.e. with the use of nous referring to weaker emotions.

One way to test this assumption is through a scrutiny of literary texts translated to and from the respective languages. Since emotions play a significant role in literature, an analysis based on a corpus of literary texts – which contain many instances of linguistic items both referring to emotions and expressing them – seems a good choice of data for comparisons. The data used in the study come from the Swedish-Polish Parallel Corpus of contemporary literary texts. The corpus contains 1,439,911 tokens of Swedish originals (14 pieces of literary fiction), and 1,227,761 tokens of their Polish translations. Since at the moment the corpus does not contain parallel data in the other direction, the hypothesis can only by tested partially.

The study focused on the nouns from the semantic field of Swedish *skräck* and Polish *strach* ('fear') and other nouns in both languages denoting related emotions. Only these items which occur more than 5 times in our corpus were selected for further analysis.

Semantic similarities and differences between the nouns denoting the emotion of *skräck* and *strach* in both languages were examined with the help of established monolingual dictionaries in both languages (see References). The analysis revealed that the semantic component of being strong or weak is one of the main differentiating features of these nouns in both languages. This may serve as a point of departure for an approximate ordering of the analysed lexical units according to the 'strong'/'weak' parameter.

The next step in our analysis involved examining how the emotions from the semantic field of ~'fear' were translated from Swedish to Polish and how the translation equivalents in both languages were distributed along the 'strong'/'weak' scale. The frequencies and the distribution of the analysed nouns in the Swedish source text subcorpus are summarised in Table 1. Table 2 captures the same information for the Polish nouns, but only those which were translations of the analysed Swedish items (and not adjectives, verbs or adverbs).

oro 'concern'[1]	**fruktan** 'concern/ anxiety'	**ängslan** 'anxiety/ apprehension'	**ångest** 'apprehension'	**rädsla** 'fear/ apprehension'	**skräck** 'fear/ dread'	**förfäran** 'horror/ terror'	**fasa** 'trepi- dation'	**panik** 'panic'	**total**
120	18	18	35	106	100	10	28	55	490
24%	4%	4%	7%	22%	20%	2%	6%	11%	100%

Table 1. The frequencies and distributions of the Swedish nouns

niepokój 'anxiety'	**obawa** 'apprehension'	**lęk** 'fear/aprehension'	**strach** 'fear/dread'	**przerażenie** 'horror/terror'	**trwoga** 'trepidation'	**panika** 'panic'	**other**	**total**
89	19	73	144	37	3	41	84	490
18%	4%	15%	29%	8%	1%	8%	17%	100%

Table 2. The frequencies and distributions of the Polish nouns in parallel sentences

The analysis indicates that Swedish texts show the tendency of denoting ~'fear' with three nouns *oro*, *rädsla* and *skräck*, which occur with similar frequencies and together represent 66% of this semantic field in the analysed texts. The Polish texts, on the other hand, rely to a much greater extent on one noun only, *strach* which is the prototypical category of the Polish field. This, however, may be a result of a translation strategy of individual translators rather than reflect differences in the expression of emotionality in both languages and cultures.

Table 3 demonstrates how the individual Swedish nouns were translated into Polish.

	niepokój	**obawa**	**lęk**	**strach**	**przerażenie**	**trwoga**	**panika**	**other**
panik				7%	5%		73%	15%
fasan			11%	11%	50%	4%		27%
förfäran				10%	60%			30%
skräck		2%	9%	65%	12%			12%
rädsla		8%	20%	58%	2%		1%	11%
fruktan		17%	28%	33%		6%		16%
ångest	11%	3%	51%	9%	3%	3%		21%
ängslan	50%		44%					6%
oro	63%	3%	8%					26%

Table 3. Nominal equivalents of the Swedish nouns in parallel sentences

The analysis of the Polish translation equivalents of the Swedish nouns referring to ~'fear' does not corroborate the initial hypothesis. For example, one of the weak Swedish nouns *ångest* – most frequently translated with a relatively weak Polish noun *lęk* (51%) – is also rendered through weaker and stronger equivalents with almost equal frequencies (14% and 15% respectively). On the other hand, the stronger Swedish nouns *förfäran* and *fasa*, which are most frequently translated by the Polish strong noun *przerażenie* (60% and 50% respectively) are also rendered by weaker nouns: *strach* (the central category, 10 and 11% respectively), and *lęk* (11%).

Thus, for the moment the hypothesis concerning the differences in the expression of emotionality in the two languages and cultures has not been confirmed. However, this study is but the first step in this endeavour. An analysis of larger datasets, other items referring to emotions (other parts of speech and other emotions) as well as an examination of the Polish original texts and their Swedish translations, planned as the next steps in the project, may shed a new light on this issue.

[1] All English glosses give only approximate meanings of individual items.

References

Athanasiadou, A. & Tabakowska, E. (eds). (1998). *Speaking of Emotions*. Berlin & New York: Mouton de Gruyter.

Coulter, J. (1986). Affects and social context: Emotion definition as a social task. In R. Harre (ed.) *The Social Construction of Emotions*. Oxford: Basil Blackwell, 120-134.

Daun, Å. (1998). *Svensk mentalitet*. Stockholm: Nordstedts Akademiska Förlag.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion* 6, 169-200.

Gruszczyńska, E. (2001). *Linguistic Images of Emotions in Translation from Polish into Swedish*. Stockholm: Elanders Gotab.

Harkins, J. & Wierzbicka, A. (eds). (2001). *Emotions in Crosslinguistic Perspective*. Berlin: Mouton de Gruyter.

Hurtado de Mendoza, A. (2008). The problem of translation in cross-cultural research on emotion concepts (commentary on Choi & Han). *International Journal for Dialogical Science* 3(1), 241-248.

Lewandowska-Tomaszczyk, B. & Wilson, P. (2013). English *fear* and Polish *strach* in contrast: GRID approach and Cognitive Corpus Linguistic Methodology. In J. Fontaine, K. R. Scherer & C. Soriano (eds). *Components of Emotional Meaning: A Sourcebook*. Oxford: Oxford University Press.

Lüdtke, U. M. (ed.) (2015). *Emotion in Language. Theory – Research – Application*. Amsterdam: John Benjamins Publishing Company.

Ortony, A., Clore, G. & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Plutchik, R. (1994). *The Psychology and Biology of Emotion*. New York: Harper Collins.

Shields, K. & Clarke, M. (eds). (2011). *Translating Emotion*. Frankfurt: Peter Lang.

Wierzbicka, A. (1988). Emotions across cultures: similarities and differences – a rejoinder to Konstantin Kolenda. *American Antropologist* 90 (4), 982-983.

Wierzbicka, A. (1990). The semantics of emotions: fear and its relatives. *Australian Journal of Linguistics* 10(2), 133-138.

Wierzbicka, A. (1992). *Semantics, Culture, and Cognition. Universal Human Concepts in Culture-specific Configurations*. New York & Oxford: Oxford University Press.

Wierzbicka, A. (1994). Emotion, language, and cultural scripts. In S. Kitayama & H. R. Markus (eds). *Emotion and culture: Empirical studies of mutual influence*. Washington: American Psychological Association.

Dictionaries

SOB Svensk Ordbok, (1990): Esselte Ordbok.

SAOB Svenska Akademiens Ordbok (Internet version) (1997). Lund Göteborg http://g3.spraakdata.gu.se/saob/.

Szymczak, M. (ed.) (1992): *Słownik języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Grammatical metaphor in translation: A corpus-based investigation

Arndt Heilmann, Tatiana Serbina, Bastian Lorenz, Stella Neumann
RWTH Aachen University
arndt.heilmann@ifaar.rwth-aachen.de, tatiana.serbina@ifaar.rwth-aachen.de,
bastian.lorenz@rwth-aachen.de, stella.neumann@ifaar.rwth-aachen.de

Grammatical metaphor is a linguistic phenomenon that allows speakers to express an event on different levels of complexity. The simple expression of an event is a clause where all of the grammatical functions, e.g. the subject, map congruently onto semantic roles, such as the agent, and are perfectly recoverable for the reader. In contrast, the same event can be expressed in a more complex way by means of a nominalization: in this case, the information is compressed and potentially left implicit (Halliday & Matthiessen 2013). The phenomenon of grammatical metaphor is illustrated in (1) below. Example (1a) shows the grammatically metaphorical realization of an event, whereas example (1b) is a possible congruent realization of the same event.

(1) a. *Grammatically metaphorical version: The creation of complex objects ... (EO POPSCI 004s106)*
b. *Congruent version: Builders create complex objects.*

Example (1b) expresses an event congruently with *builders* as the subject mapping onto the semantic role of the agent and *complex objects* as a direct object, mapping onto the role of a goal. In (1a), the same event is expressed using a grammatical metaphor: an event is aligned with a noun phrase with the nominalization *creation* functioning as its head. In this case, the actual agent of the event is left implicit. One of the advantages of using grammatically metaphorical nominal expressions is their flexibility in terms of a potential modification (e.g. *a difficult creation of complex objects*) and a possibility of being combined with further clauses (e.g. to function as the subject of the sentence).

Steiner (2001) suggests that translation of nominalized, i.e. grammatically metaphorical, expressions may involve the process of understanding during which the translator 'unpacks' the nominalized event to its congruent form before translating it. Due to potentially higher effort required to translate grammatically metaphorical stretches of text, it is furthermore assumed that the level of grammatical metaphoricity is likely to be reduced in the process of translation (Steiner 2001; Hansen-Schirra & Steiner 2012), thus leading to translation shifts towards verbal expressions. At the same time, Hansen-Schirra & Steiner (2012) do not exclude a possibility that entrenched expressions are translated directly without the initial 'unpacking', resulting in equally or even more condensed nominal expressions in the corresponding translations. Such translation behavior is also predicted by the literal translation hypothesis (Tirkkonen-Condit 2005), according to which the literal translation is typically translator's first choice.

Previous research (Alves et al. 2010) has investigated (de-)metaphorization processes in several registers of the cross-linguistic CroCo corpus taking into account part-of-speech changes and the alignment between grammatical functions for the language pair English-German. Alves et al. found that English translations were less nominal than their corresponding German originals. Furthermore, several clauses in the English translations are frequently aligned to German nominalizations. These translation shifts could be interpreted as indications of reducing grammatical metaphoricity in the translation direction German-English. Moreover, in the opposite translation direction, studies have shown that among part-of-speech changes, shifts from verbs to other parts of speech are particularly common (Alves et al. 2010; Serbina et al. 2017), potentially increasing the level of metaphoricity. Thus, these results suggest that the type of changes could differ depending on the translation direction. The present study aims at further testing this assumption considering the partly contradictory hypotheses by Steiner (2001) and Tirkkonen-Condit (2005) discussed above.

Similar to Alves et al. (2010), we used the parallel English-German CroCo corpus (Hansen-Schirraet al. 2012) but concentrated on the register of popular-scientific writing and on a specific complex nominal construction. English and German originals and translations were queried for nouns embedded into the English genitive *of*-

construction and its corresponding German construction. The sentences containing these constructions as well as the aligned units were then annotated in terms of grammatical metaphor. For instance, in (2) the noun phrase *the development of a new organism* was translated as the noun phrase *die Entwicklung eines neuen Organismus* ('the development of a new organism'), thus this alignment pair was annotated as an instance of re-metaphorization.

(2) EO: *This dynamic sequence of events with its changing pattems* [sic!] *of gene activities during cell reproduction is called the genetic program and it directs the development of a new organism*. (EO POPSCI 008s85)
GTRANS: *Diese dynamische Abfolge von Ereignissen mit ihrem wechselnden Muster von Genaktivitäten während der Fortpflanzung nennt man das genetische Programm, das die Entwicklung eines neuen Organismus steuert.*

The initial quantitative analysis indicates that translators typically do not introduce translation shifts. However, it seems to be the case that demetaphorization is slightly more common in the direction from English to German. Therefore, these results corroborate the assumption of a direct translation not involving the process of 'unpacking' as well as the literal translation hypothesis. However, example (2) also illustrates one of the problems connected to a quantitative analysis of grammatical metaphors in translation: if we concentrate on the noun phrases that were automatically extracted from the corpus, then the example is classified as remetaphorization. However, if we consider the whole sentences in the original and the corresponding translation, (2) could be considered as an example of metaphorization, since the analyzed noun phrase is embedded in a coordinated clause, whereas its translation is part of the postmodification of the noun *Programm* ('program').

In the next step, we will test the observed tendencies by performing a multifactorial regression model. Moreover, the analyzed sentences will be investigated in more detail taking into account their internal composition, i.e. phrase structures and parts-of-speech. We also show differences of (re-)metaphorization behaviour, depending on the subject areas (e.g. philosophical texts compared to other subject matters). The analysis is expected to present a more comprehensive picture of grammatical metaphor in translation.

References

Alves, F., Pagano, A., Neumann, S., Steiner, E. & Hansen-Schirra, S. (2010). Translation units and grammatical shifts: towards an integration of product and process-based translation research. *Translation and cognition*, 109-142.

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2013). *An Introduction to Functional Grammar.* London: Routledge.

Hansen-Schirra, S., Neumann, S. & Steiner, E. (eds). (2012). *Cross-Linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German.* Berlin: De Gruyter.

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

Serbina, T., Hintzen, S., Niemietz, P. & Neumann, S. (2017). Changes of word class during the translation process: Insights from a combined analysis of keystroke logging and eye-tracking data. In S. Hansen-Schirra, O. Czulo & S. Hoffmann (eds). *Empirical modelling of translation and interpreting*. Berlin: LangSci Press, 177-208.

Steiner, E. (2001). Translations Englih-German: Investigating the Relative Importance of Systemic Contrasts and the Text-Type "Translation". *SPRIKreports: Reports of the Project Languages in Contrast* 7, 1-49.

Tirkkonen-Condit, S. (2005). The Monitor Model Revisited: Evidence from Process Research. *Meta* 50 (2), 405-414.

The conditional in English and French: A comparable-parallel corpus analysis

Daniel Henkel
Université Paris 8 Vincennes St-Denis TransCrit EA1569
daniel.henkel@univ-paris8.fr

Introduction

This study is part of a long-term project in three phases, the aim of which is to determine quantitatively, using a series of syntactic and lexical indicators, first of all, what statistically significant differences can be demonstrated when English-translated-from-French (EtrF) is compared with “original” or *ex nihilo* English (En0), and when French-translated-from-English (FtrE) is compared with “original” or *ex nihilo* French (Fr0), secondly why such differences exist, and thirdly what recommendations may be made to achieve greater similarity between translations and the target language. The analysis presented herein belongs to the first phase of the study and focuses on conditional mood, in keeping with Lyons' (1995) definition of “mood” as “that category which results (…) from the grammaticalization of subjective modality and other kinds of subjective meaning” (Lyons 1995: 179), as represented by the French morphological conditional and its closest equivalents in English, i.e. WOULD and the other preterit modal auxiliary verbs (COULD, SHOULD, MIGHT, OUGHT). At the same time, it will be shown methodologically how the quantitative analysis of comparable corpora (i.e. corpora in different languages composed of untranslated texts sharing common characteristics) and parallel corpora (i.e. corpora consisting of source- and target-texts) can yield greater insight than either approach on its own.

Methods

Observations were made using an 8-million-word corpus consisting of four 2-million-word subcorpora:

- 20 public-domain works in original English by 20 different authors,
- the translations of these into French,
- 20 public-domain works in original French by 20 different authors,
- the translations of these into English,

as a basis for three different quantitative comparisons:

- between original English (En0) and original French (Fr0),
- between English-translated-from-French (EtrF) and original English (En0), and between French-translated-from-English (FtrE) and original French (Fr0).
- between source- and target-texts.

Translated works of the late 19th-early 20th centuries were collected according to the availability of both the original and its translation in electronic format, using common inclusion criteria based on size and date of publication, so as to obtain two pairs of corpora comparable in terms of size, stylistic diversity and period. All 80 original and translated texts were tagged by POS and lemma with TreeTagger, and analyzed in TextSTAT using regular expressions targeting preterit modals in English and the conditional in French. The results of each query were converted to normed frequencies expressed as occurrences per 1000 words (Freq./1k). Intra-linguistically the Wilcoxon-Mann-Whitney rank-sum test was used to compare authors and translators and thus to determine whether the language produced by translators displays the same overall characteristics as the language produced by authors in their own language. Finally, the frequencies of these corresponding grammatical structures in the source- and target-texts were evaluated using Spearman's correlation to test for possible inter-linguistic influences.

Results

On the whole, the frequency of WOULD and other preterit modals was found to be much higher in En0 (median frequency 8.31/1k words) than that of the conditional in Fr0 (median 2.97/1k), while EtrF (median 5.87/1k) and FtrE (median 4.99/1k) occupy an intermediate zone as shown in Figure 1.

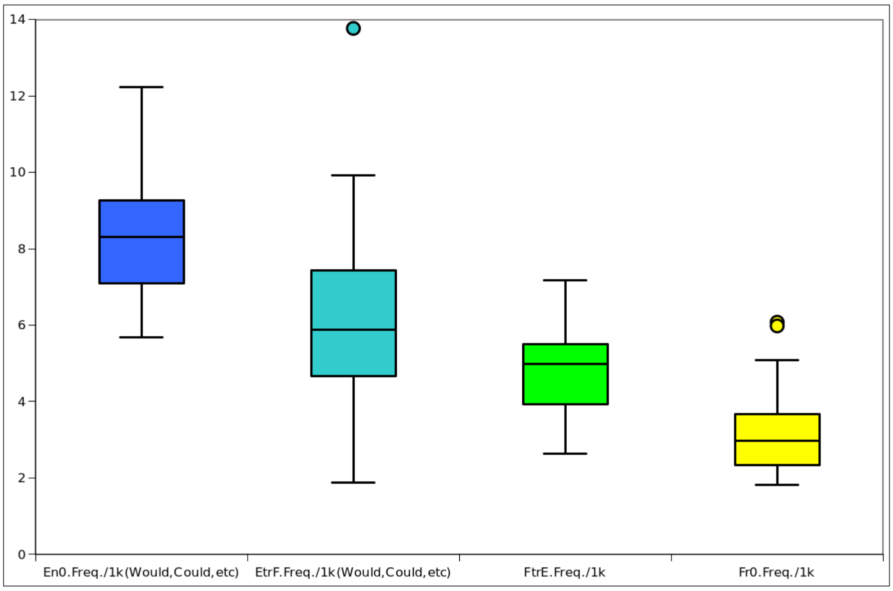


Figure 1. Frequency of conditional forms in En0, EtrF, FtrE and Fr0

Moreover, the Wilcoxon-Mann-Whitney test reveals a statistically significant disparity in each translated/*ex nihilo* pair:

- for English, EtrF vs. En0 U=95, p=0.004 (n1=n2=20)
- for French, FtrE vs. Fr0 U=327, p<0.001 (n1=n2=20)

These findings thus provide strong evidence of reciprocal interlinguistic interferences that both induce translators to overuse the conditional when translating into French, and to neglect conditional modals when translating into English.

It should further be emphasized that the choice of WOULD and the French conditional as indicators to compare translated and *ex nihilo* texts is not meant to imply that these two forms are strictly equivalent to one another. It is well-known that WOULD, in particular, in addition to its counter-factual meaning, has other semantic interpretations which are not shared with the French conditional, most notably its use in expressing iterative aspect as evidenced in the following examples:

1. *And yet he would always wind up by muttering that no sister of his should ever have accepted such a situation.* (A. Conan Doyle, The Adventures of Sherlock Holmes) → *Mais il en venait toujours à répéter ce qu'il avait dit en premier lieu : que, s'il avait eu une sœur, il ne lui aurait jamais permis d'accepter une situation comme celle-là.*

2. *He would often spend a whole day settling and resettling in their cases the various stones that he had collected …* (O. Wilde, The Picture of Dorian Gray) → *Il passait souvent des journées entières, rangeant et dérangeant dans leurs boîtes les pierres variées qu'il avait réunies …*

Notwithstanding such divergences, Spearman's coefficient indicates a near-perfect (ρ=0.91, p<0.001) correlation between the frequency of the conditional modals in the EtrF target-texts and that of the French conditional in the Fr0 source-texts, which is even stronger (ρ=0.97, p<0.001) for WOULD itself. Likewise, a statistically significant (p<0.001), moderately strong correlation (ρ=0.71 overall, ρ=0.68 for WOULD alone) exists in the other direction between the frequency of the conditional in the FtrE target-texts and that of the conditional modals in the En0 sources, which may be a consequence of the greater range of semantic interpretations to be found among the English modals. These correlations are illustrated in the form of scatterplots in Figures 2a-b.

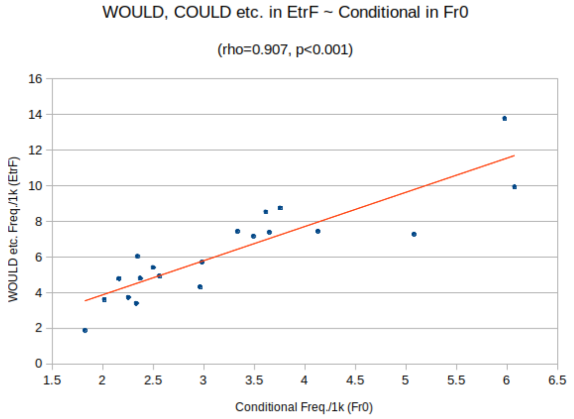


Figure 2a. Frequency of conditional modals in EtrF compared to Fr0 conditional.

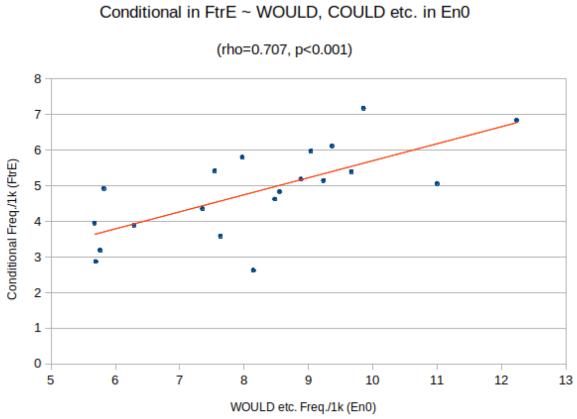


Figure 2b. Frequency of FtrE conditional compared to conditional modals in En0.

Conclusion

Much more could be said about the translation of the conditional between English and French if space allowed. This brief synopsis nonetheless demonstrates how the quantitative analysis of comparable-parallel corpora can be used both to identify disparities between target-texts and the target-language as represented in an *ex nihilo* corpus, and to assess the influence of the source-texts on the target-texts. In this case, both English-translated-from-French and French-translated-from-English were shown to be significantly different from *ex nihilo* English and French in their use of the conditional. In the process of translation, the correlation between the Fr0 conditional and the corresponding modals in EtrF is nearly perfect, while the statistically significant yet somewhat weaker correlation between En0 source- and FtrE target-texts may well be a reflexion of the greater semantic diversity of the English modals.

References

Bernardini, S. (2011). Monolingual comparable corpora and parallel corpora in the search for features of translated language. *Synaps* 26, 2-13.

Hu, K. (2016). *Introducing corpus-based translation studies*. New York: Springer.

Hüning, M. TextSTAT 2.9c. (2000/2014). Niederländische Philologie. Freie Universität Berlin, http://neon.niederlandistik.fu-berlin.de/en/textstat/.

Kruger, A., Wallmach, K. & Munday, J. (eds). (2011). *Corpus-based translation studies: Research and applications*. London: Bloomsbury Publishing.

Loock, R. (2016). *La Traductologie de corpus*. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge: Cambridge University Press.

Olohan, M. (2002). Comparable corpora in translation research: Overview of recent analyses using the translational English corpus. In *LREC Language Resources in Translation Work and Research Workshop Proceedings*, 5-9.

R Development Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rohde, D. L. (2000). Extracting Syntax Statistics from Large Corpora of Written English.

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

Zanettin, F. (2013). Corpus methods for descriptive translation studies. *Procedia-Social and Behavioral Sciences*, 95, 20-32.

Exploring new approaches to the corpus-based contrastive study of hedging strategies in spoken language

Stine Hulleberg Johansen
University of Oslo
s.h.johansen@ilos.uio.no

In recent years corpus linguistics and pragmatics have begun exploring their common ground. However, this has not been altogether straightforward, mainly because "core features of pragmatics studies (...) are harder to catch with corpus methodology than lexical or morpho-syntactic features" (Taavitsainen & Jucker 2015: 12). One such core feature is that of hedging. Hedging strategies can take almost any linguistic (or paralinguistic) form and is not an inherent property of words or phrases (Stenström 1994). Thus identifying hedging strategies is challenging without a pragmatically annotated corpus.

Consequently, in the absence of pragmatically annotated corpora, two main ways of studying pragmatic phenomena through corpus linguistic methodologies have emerged. One is the form-to-function approach that starts from pre-defined lexical words or constructions whose potential pragmatic uses are examined (Aijmer & Rühlemann 2015). The second is the function-to-form approach, which starts from a language function and investigates the forms used to perform that function.

Although these approaches are presented as equally relevant in the literature, researchers show a clear preference for the former. This is not surprising as a major challenge with the latter approach is that the function cannot be retrieved, only surface forms orbiting it can be used to identify the function in the corpus. This raises the question of whether the function-to-form approach is a realistic methodological alternative. Moreover, there is a need to understand how this approach actually manifests itself and how it can be applied in corpus-based contrastive studies of pragmatic phenomena to capture cross-linguistic variation.

The present study explores one potential application of the function-to-form approach. By searching for certain characteristics of situations where hedging strategies tend to occur, the study aims to retrieve various realisations of hedging strategies. More specifically, by using the conventionalised direct non-performative refusal strategy *no* (English) and the corresponding *nei* (Norwegian) (Beebe et al. 1990) as well as the conjunction *but* (English) and *men* (Norwegian) signalling contrast or denial of expectation (Blakemore 1989) as framing devices, the aim is to identify co-occurring hedging strategies in these face-threatening situations. This leads to the following research questions:

RQ1: How can we identify framing devices for extracting pragmatic functions from corpora?
A, conventionalised realisations of speech acts
B, explicit signals of contradiction/contrast
RQ2: Will this application of the function-to-form approach work across languages (Norwegian and English) allowing for a comparison of two or more languages?

The choice of *no/nei* and *but/men* as tools in retrieving hedging strategies is rooted in pragmatic research on speech acts and politeness. Politeness is considered a primary motivation for using hedging strategies in conversations (Markkanen & Schröder 1997). Moreover, refusals have proven to be intrinsically face-threatening across various cultures (Demirkol 2016). Thus, it is likely that heding strategies will co-occur with refusals as a way of softening the blow. Similarly, saying something that contradicts or is in contrast to what has previously been said can also threaten the hearer's positive face (Brown & Levinson 1987). Even contradicting oneself is considered threatening to the speaker's positive face. Thus identifying conventionalised realisations of refusals or contradictions can be instrumental in locating hedging strategies within a corpus.

In this study, direct refusals will be retrieved from four spoken corpora: BNC2014, Nordic Dialect Corpus (NDC), Norwegian Speech Corpus (NoTa) and the BigBrother corpus (BB). Only the conversational part of the NDC and NoTa will be used to make the data more comparable. There are no bidirectional or directly comparable corpora of spoken Norwegian and English, thus the corpora in this study are chosen based on their degree of comparability and their availability. This allows for a comparison of the results between the two languages.

In the study, 150 random instances of *nei* and *no* and 150 random instances of *men* and *but* in the respective languages were chosen from the corpora. Table 1 shows the number of *nei/no* used as refusals and the number of contrastive uses of *men/but* among the 150 instances in each language. The columns to the right show how many of these occurrences that co-occur with hedging strategies.

	nei/no	***men/but***	***nei/no* with hedging**	***men/but with* hedging**
Norwegian corpora (NDC, NoTa, BB)	31	135	12	93
English corpus (BNC2014)	85	149	28	84
Sum	116	284	40	177

Table 1. Number of occurrences of *nei/no* and *men/but* and their co-occurring hedging strategies in the four corpora

Although 38.7 % and 32.9 % of the occurrences of *nei* and *no* were hedged, there were only 116 instances of the total of 300 *nei* and *no* being used as refusals in the data. This indicates that in order to use *nei* and *no* as framing devices, one would have to manually process a great deal of data to retrieve a sensible amount and variety of hedging strategies. In contrast, *men* and *but* showed more promising numbers with 284 relevant instances and 68.9 % and 56.4 % co-occurring with hedging strategies respectively. Example 1 below illustrates a typical example of hedging strategies (italicised) co-occurring with *but* in the English dataset.

Example 1 from BNC2014

S0598: you 're nearly an adult —ANONnameF
S0596: I am an adult
(…)
S0596: I am an adult I can vote
S0598: yeah **but** *what I mean is like* you can still say that you 're a teenager *though* cos eighteen

Preliminary results suggest that conventionalised realisations of face-threatening speech acts and the like can be used to identify other language functions in a corpus. However, the choice of framing device must be carefully selected and tested. In this study, both *no* and *nei* and *but* and *men* co-occurred with hedging strategies, but *but* and *men* returned the highest number of hedging strategies and the greatest variation between realisations. However, more data need to be analysed to confirm this. Furthermore, using this approach to study how a particular function, with potentially indefinite realisations, is realised can be fruitful in the absence of pragmatically annotated corpora, particularly to capture linguistic variation across languages.

References

Aijmer, K. & Rühlemann, C. (2015). *Corpus pragmatics: a handbook*. Cambridge: Cambridge University Press.

Beebe, L. M., Takahashi, T. & Uliss-Weltz, R. (1990). Pragmatic Transfer in ESL Refusals. In R. C. Scarcella, E. S. Andersen & S. D. Krashen (eds). *Developing communicative competence in a second language*. New York: Newbury House Publishers, 55-73.

BigBrother-korpuset, Tekstlaboratoriet, ILN, Universitetet i Oslo. http://www.tekstlab.uio.no/nota/bigbrother/.

Blakemore, D. (1989). Denial and contrast: a relevance theoretic analysis of *but*. *Linguistics and Philosophy*, 12, 15-37.

Brown, P. & Levinson, S. C. (1987). *Politeness: some universals in language usage*. Cambridge: Cambridge University Press.

Demirkol, T. (2016). How Do We Say 'No' in English? *Procedia — Social and Behavioral Sciences*, 232, 792-799. doi:10.1016/j.sbspro.2016.10.107.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A. & Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus — an Advanced Research Tool. In K. Jokinen & E. Bick (eds). *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA2009), May 14-16, Odense, Denmark*. Northern European Association for Language Technology (NEALT).

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.

Markkanen, R. & Schröder, H. (1997). *Hedging and discourse: approaches to the analysis of a pragmatic phenomenon in academic texts* (Vol. 24). Berlin: Walter de Gruyter.

Norsk talespråkskorpus — Oslodelen, Tekstlaboratoriet, ILN, Universitetet i Oslo. http://www.tekstlab.uio.no/nota/oslo/index.html.

Stenström, A.-B. (1994). *An introduction to spoken interaction*. London: Longman.

Taavitsainen, I. & Jucker, A. H. (2015). Twenty years of historical pragmatics: Origins, developments and changing thought styles. *Journal of Historical Pragmatics*, 16(1), 1-24.

Detecting traces of constrained communication: A corpus-driven approach to mapping of the intersection between learner language and translated language

Ilmari Ivaska, Silvia Bernardini, Adriano Ferraresi
University of Bologna
ilmari.ivaska@unibo.it, silvia.bernardini@unibo.it, adriano.ferraresi@unibo.it

This contribution explores possible linguistic traces of constrained communication across non-native and translated language. Both varieties (L2 Original, or L2O, and L1 Translated, or L1T), have been suggested to diverge from first/non-translated language (L1 Original, or L1O). We seek to map this intersection in a corpus-driven manner, detecting commonalities in the way L2O and L1T diverge from L1O. Our research questions are: 1) Are there any linguistic features that distinguish both L2O and L1T English from L1O? 2) How can the presence of such common features, if any, be interpreted in the light of the constrained communication hypothesis?

Theory and earlier research

Corpus studies of non-native language use and translation share a common interest in contrasting these linguistic varieties to the baseline of "native" production, and in both disciplines a substantial body of research has investigated typical features of L2O/L1T when compared to L1O (Granger 2015; Xiao & Hu 2015). It has also been proposed that instances of language use where more than one language is inherently present (Lanstyák & Heltai 2012; Kruger & van Rooy 2016) may have typical features in common, even though situational constraints are obviously of a very different nature (Kolehmainen et al. 2014). By bringing these two types of constrained communication together within a single methodological framework, it might thus be possible to attain higher-order, more powerful generalizations than has hitherto been the case.

Material and methods

In the present paper, we implement a quantitative bottom-up approach to the identification of linguistic features distinguishing L2O and L1T from L1O. We then analyse the typical use of one such feature, so as to better understand the nature of the divergence.

We use a composite corpus drawn from a variety of existing resources, with the addition of texts collected ad hoc to improve genre coverage and comparability across varieties. The existing corpora are the British National Corpus (BNC 2007), the Corrected and Structured EuroParl corpus (CoStEP; Graën et al. 2014), and the highermost level of EF Cambridge Open Language Database (EFCamDat; Geertzen et al. 2013). Our final dataset comprises 900 texts (or 314,811 tokens), divided equally between three genres (EuroParl, news, and touristguides) and three varieties (L2O, L1T, and L1O). Unfortunately, the sources do not include enough L1-/source-language-specific data to allow comparisons taking this factor into account.

We annotated the data using the UDPipe parser (Straka & Straková 2017), and extracted typical features that distinguish the studied varieties using key structure analysis (Ivaska 2015; Ivaska & Siitonen 2017). Similarly to keyword analysis, key structure analysis uses n-gram frequencies (here, unigrams of parts-of-speech, morphological features, and syntactic dependencies normalized over 1,000 words per text) and random forests (e.g. Tagliamonte & Baayen 2012) to zoom in on grammatical features over- and under-represented in a given set of texts. We consider the 10 best unigram predictors of the difference between L2O / L1O and L1T / L1O, and identify those common to both lists. We then analyse lexical, structural and cotextual variation (collocations and colligations) of one such feature.

Preliminary results

We identified multiple linguistic features whose frequency of use distinguishes both L2O and L1T from L1O: 8/10 parts-of-speech, 4/10 morphological features, and 3/10 syntactic features are common to both lists of best predictors. To focus on features in which the difference is not exclusively genre-specific, we excluded the features in which the difference was not portrayed in at least two of the three genres in both comparisons. The most consistent difference could be seen in syntactic dependencies: in both news texts and tourist guides, syntactic root elements were more numerous in L1T and L2O than in L1O (see Figure 1). Interestingly, the same general pattern applies to both genres, whereby translations occupy the middle position between native and non-native texts.

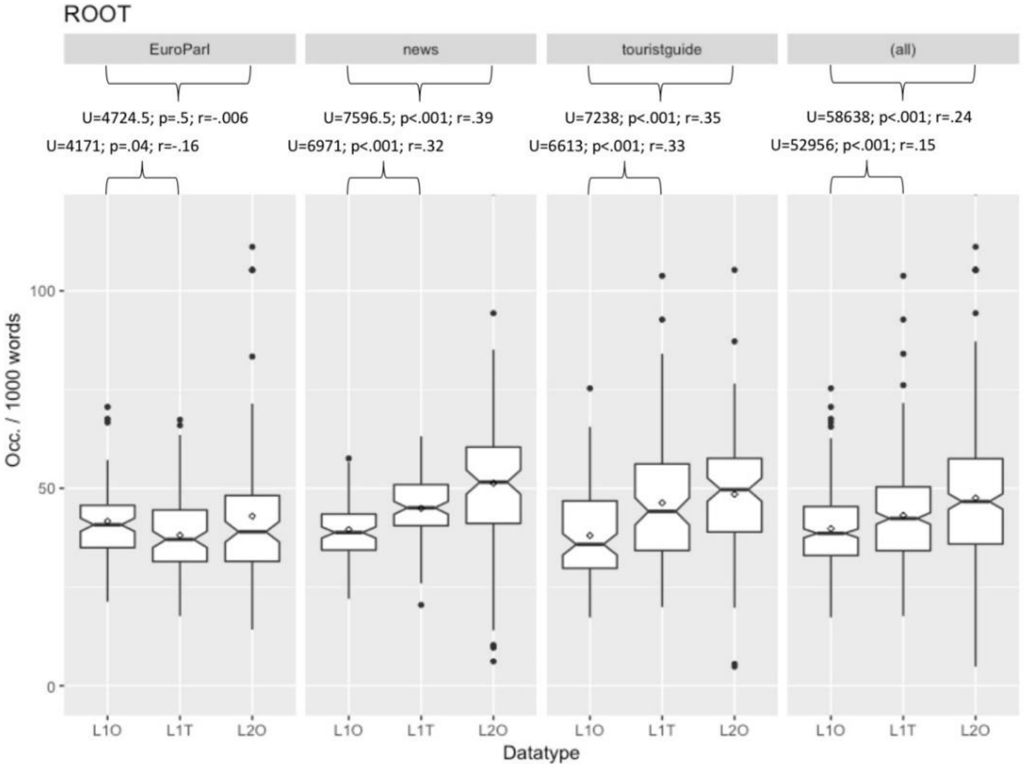


Figure 1. Relative frequencies of the syntactic dependency root in different genres

When focusing on inner and cotextual variation of the root element, the main difference between L2O/L1T and L1O is sentence length: sentences are generally longer in L1O than in L2O/L1T. Other relevant aspects are parts-of-speech, syntactic dependencies and lemmas assigned to the word that precedes the root element. In particular, the typical features preceding the root element (auxiliary *be* and nominal subject) cover a larger portion of the observations in L1T/L2O than L1O. Together with the higher frequency of root elements, sentence length and auxiliary be seem to point to greater simplicity and less syntactic variation in L1T/L2O in comparison to L1O.

Discussion and conclusion

Our study provides support for the constrained language hypothesis, as several linguistic features that distinguish both L2O and L1T texts from L2O texts, in two different genres, could be identified in a corpus-driven manner. Among the predictors of translated-ness/non-nativeness, we singled out for further analysis the syntactic root element, and observed that its lower frequency may tentatively be related to greater syntactic variation and complexity in sentence structure in non-constrained texts. The interpretation is in line with earlier observations by Kruger & van Rooy (2016) regarding the higher normativity and prototypicality of the constrained varieties, potentially related to their higher cognitive demands (Rohdenburg 1996).

Further study is necessary to confirm this interpretation of results, to investigate other linguistic features typical of constrained language use, and to explore language-pair specific datasets, which might highlight cross-linguistic influences. Furthermore, we plan to apply this method to other languages: the language-agnostic research design and methodological pipeline applied allows us to replicate this study in any language with sufficient data and a parser following the universal annotation scheme, thus enabling further exploration towards the bottom-up detection of universal tendencies of constrained communication.

References

BNC 2007 = The British National Corpus. (2007). Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. BNC XML edition. http://www.natcorp.ox.ac.uk/.

Geertzen, J., Alexopoulou, T. & Korhonen, A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Selected Proceedings of the 31st Second Language Research Forum (SLRF)*. MA: Cascadilla Press.

Graën, J., Dolores B. & Volk, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. In *Konvens 2014*, 8-10 October 2014, Hildesheim, Germany, 222-227.

Granger, S. (2015). Contrastive Interlanguage Analysis: A Reappraisal. *International Journal of Learner Corpus Research* 1(1), 7-24.

Ivaska, I. (2015). Longitudinal Changes in Academic Learner Finnish: A Key Structure Analysis. *International Journal of Learner Corpus Research* 1(2), 210-241.

Ivaska, I. & Siitonen, K. (2017). Learner Language Morphology as a Window to Crosslinguistic Influences: A Key Structure Analysis. *Nordic Journal of Linguistics* 40(2), 225-53.

Kolehmainen, L., Meriläinen, L. & Riionheimo, H. (2014). Interlingual Reduction: Evidence from Language Contacts, Translation and Second Language Acquisition. In H. Paulasto, L. Meriläinen, H. Riionheimo & M. Kok (eds). *Language Contacts at the Crossroads of Disciplines*. Cambridge: Cambridge Scholars Publishing, 3-32.

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.

Lanstyák, I. & Heltai, P. (2012). Universals in Language Contact and Translation. *Across Languages and Cultures* 13(1), 99-121.

Rohdenburg, G. (1996). Cognitive Complexity and Increased Grammatical Explicitness in English. *Cognitive Linguistics* 7(2), 149-182.

Straka, M. & Straková, J. (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*. Vancouver, Canada: Association for Computational Linguistics, 88-99.

Tagliamonte, S. A. & Harald Baayen, R. (2012). Models, Forests and Trees of York English: Was/Were Variation as a Case Study for Statistical Practice. *Language Variation and Change* 24, 135-178.

Xiao, R. & Xiaynao, H. (2015). *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*. Berlin: Springer-Verlag Berlin Heidelberg.

Cross-linguistic register analysis in specialised discourse. A corpus-based investigation of denominal adjectives in LSP: the examples of medicine and earth sciences

Tiffany Jandrain
F.R.S.-FNRS & University of Mons
tiffany.jandrain@umons.ac.be

Noun phrases appear to be linguistic features that are problematic for translation students when they translate specialised texts from English into French. Many specialised noun phrases in French may indeed be constructed in two ways, i.e. the noun of the phrase is modified by a relational adjective or a prepositional phrase complement. Therefore, the translator has to choose between these options depending on the languages and the context of communication involved, i.e. the communicative event, including the convention which states that a linguistic utterance is appropriate or not to a specific language use, according to Systemic Functional Linguistics theory (Hatim & Mason 1990). For instance, in medical discourse, a non-expert uses *cancer du sein* while an expert tends to use this term or the relational adjective form *cancer mammaire*, which sounds more technical, depending on the message receivers and the communicational situation (Maniez 2009). These sociolinguistic factors seem therefore to play a major role in this choice (*ibid.*), especially for LSP ("Language for Specific Purposes") texts given that domain-specific languages can be considered "contextual-functional varieties of the ordinary language" and thus vary according communication function and context (Garzone 2006 in Pignataro 2012: 128).

Based on Maniez's study, this paper aims to analyse the influence of register on this choice. The functional approach defines register as a variety of language considered appropriate to the communicational context in which the text occurs and thus patterning language use (Halliday & Matthiessen 2014). Indeed, as registers are determined by the communicational context and function of the text (Biber & Conrad 2009), they are unsurprisingly one of the text features that a translator may pay attention to in order to make their translation appropriate to the target audience, as stated by the functionalist theory (Nord 2006). In other words, given that a specialist of a discipline is thought to use relational adjectives in specialised contexts (and thus within specialised registers and genres, i.e. "[c]onventional forms of texts associated with particular types of social occasion" (Hatim & Mason 1990: 241), it may be interesting to analyse how denominal adjectives, which make up most relational adjectives, are used in two different specialised genres in two different discourses that presumably have common characteristics (same registers, etc.) but also display differences. This analysis is carried out in English and French since it may be interesting to have a closer look at expected similarities and differences between them given that their registers operate specifically to their systems and may thus lead to variation (Chuquet & Paillard 1987). Respective cultures may also be thought to play a role in contrastive differences in academic prose (Galtung 1981). More concretely, this contrastive study compares choices made by English and French speakers in research articles from specialised journals and research information published on websites of specialised departments and institutes in the fields of medicine and earth sciences. It also gives an overview of the use of noun pre-modifiers in English. The aim of analysing original discourse is to provide a deeper insight of specialised discourse mechanisms and thus to offer useful guidelines to translation students to translate noun phrases in French since registers influence the distribution and use of most linguistic features and variations (Biber 2010) and thus play a major role in language description.

As several previous studies on register variation have shown, corpora appear to be a useful and relevant tool to explore linguistic features influenced by registers. In fact, registers are a phenomenon of frequency: they are characterised by recurring linguistic features which are themselves shaped by the communicational context and use and become the norm of use through their repetition in that context (Neumann 2016). In other words, register studies necessarily require a quantitative analysis, which can be accomplished with the use of corpora (Giménez-Moreno & Skorczynska 2013). This first analysis will allow to see whether there is a statistically significant difference between noun phrase modification use in registers and discourses in English and French.

It will be followed by a qualitative examination of the results to draw preliminary conclusions about this use. Our hypothesis is that experts of both disciplines will use more denominal adjectives than prepositional phrase complements despite register variation.

This corpus-based investigation is in the lineage of the study of non-literary register variation, which has been overlooked by contrastive linguistics and translation studies, which may appear quite astonishing since registers crucially influence cross-linguistic contrasts (Lefer & Vogeleer 2016). It is therefore also in the lineage of answering the urgent call recently made by scholars to carry out more register analyses, especially in order to provide guidelines based on truthful examples to translation students (see for example Vandaele 2015). It also presents the methodological criteria used in this analysis to compile specialised corpora according to genres and registers, which are non-consensual notions as studies have shown (see for example Lee 2001). It eventually suggests new paths of research in cross-linguistic register analysis for linguistic and translation purposes.

References

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine & H. Narrog (eds). *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, 159-192, http://eclass.uoa.gr/ (retrieved on 02/08/2017).

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

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

Galtung, J. (1981). Structure, culture, and intellectual style: An essay comparing saxonic, teutonic, gallic and nipponic approaches. *Social Science Information* 20(6), 817-856, http://www.transcend.org/ (retrieved on 11/04/2018).

Giménez-Moreno, R. & Skorczynska, H. (2013). Corpus analysis and register variation: a field in need of an update. *Procedia — Social and Behavioral Sciences* 95, 402-408, http://www.sciencedirect.com/ (retrieved on 19/09/2017).

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar (Fourth Edition)*. Oxon & New York: Routledge.

Hatim, B. & Mason, I. (1990). *Discourse and the Translator*. London & New York: Longman.

Lee, D. Y. M. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3), 37-72, http://www.llt.msu.edu/ (retrieved on 10/10/2017).

Lefer, M.-A. & Vogeleer, S. (eds). (2016). *Register- and Genre-related Discourse Features in Contrast*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Maniez, F. (2009). L'adjectif dénominal en langue de spécialité : étude du domaine de la médecine. *Revue française de linguistique appliquée* 14(2), 117-130.

Neumann, S. (2016). Cross-linguistic register studies. In M.-A. Lefer & S. Vogeleer (eds). *Register- and Genre-related Discourse Features in Contrast*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Nord, C. (2006). Loyalty and fidelity in specialised translation. *Confluências – Revista de Tradução Científica e Técnica* 4, 29-41, http://www.web.letras.up.pt/ (retrieved on 08/02/2016).

Pignataro, C. (2012). Terminology and Interpreting in LSP Conferences: A Computer-aided vs. Empirical-based Approach. In C. J. Kellet Bidoli (ed.) *Interpreting across Genres: Multiple Research Perspectives*. Trieste: Edizioni Università di Triste, 125-140, http://www.academia.edu/ (retrieved on 14/12/2017).

Vandaele, S. (2015). La recherche traductologique dans les domaines de spécialité : un nouveau tournant. *Meta* 60(2), 209-235.

Cross-linguistic (dis)similarities in translation: process and product

Moritz Jonas Schaeffer[1], Katharina Oster[1], Jean Nitzke[1], Anke Tardel[1],
Anne-Kathrin Gros[1], Silke Gutermuth[1], Silvia Hansen-Schirra[1], Michael Carl[2,3]
Johannes Gutenberg Universität-Mainz[1], Renmin University of China[2], Copenhagen Business School[3]
mschaeffer@uni-mainz.de, osterk@uni-mainz.de, nitzke@uni-mainz.de, antardel@uni-mainz.de,
a.gros@uni-mainz.de, gutermsi@uni-mainz.de, hansenss@uni-mainz.de, m.gummiball@googlemail.com

Introduction

A number of different sources describe a phenomenon referred to as the literal or default translation hypothesis (e.g. Halverson 2015; Tirkkonen-Condit 2005; Malmkjær 2011; Schaeffer & Carl 2013; Schaeffer & Carl 2014; Schaeffer et al. 2016). Schaeffer & Carl (2013) propose that the semantic and syntactic aspects of the source text (ST) have a cross-linguistic priming effect on translators when translating into a target text (TT) and further argue that this priming effect results in more or less literal translations. In other words, how the ST is represented cognitively has an effect on the TT. These aspects have been operationalised with two measures — one which describes the semantic similarity and a second one which describes the syntactic similarity of ST and TT items. Semantic similarity is captured by word translation entropy (*HTra*) (Carl et al. 2016), i.e., the predictability of the translation of a particular ST word. The *Cross* feature (Carl et al. 2016) captures to what extent the position of ST and TT words differs. The *HCross* feature (Carl & Schaeffer 2017) describes the word order choices a translator has given a source item (*HCross* is calculated as the entropy of ST words on the basis of the *Cross* values). The literal translation hypothesis proposes that the more similar, in terms of semantics and syntax, the ST and TT items are, the more literal these translations are and the stronger the priming effect (Schaeffer & Carl 2013). It has been shown that these measures have an effect on both eye movements and typing behaviour during translation: the greater the semantic and / or syntactic similarity between ST and TT items, the less effortful is the translation and the more automatic is the process (e.g. Schaeffer et al. 2016). In other words, it is possible to predict the translation behaviour on the basis of the final TT.

The current study

We use the existing metrics (*HTra* and *HCross*) and the TPR-DB (Carl et al. 2016). The TPR-DB is a large and unique corpus which contains eye movement and keystroke data during translation and postediting. It offers over 200 features with which the data can be described. it is available under a creative commons licence (https://sites.google.com/site/centretranslationinnovation/tpr-db). For the current study, we used a subset of this database containing data from 175 translators, 619 source text words and over 80.000 target words produced during translation and postediting. The subset contains translations from English into six different languages (German, Danish, Spanish, Hindi, Japanese and Chinese). We only have limited data on translation experience, language proficiency, language dominance and typing skill, but all participants translated into their L1.

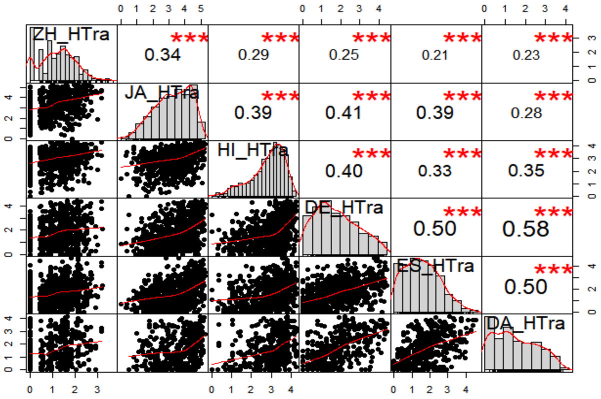


Figure 1. Semantic cross-linguistic similarity as measured by the correlation of HTra values in 6 different target languages

We show, in preliminary results, that the semantic and syntactic similarities between source and target are different for different language families (see Figure 1): In terms of semantic aspects, European languages (Danish, Spanish and German) are more similar to each other than to Asian languages (Hindi, Chinese and Japanese) and the Asian languages are also more similar to each other than to the European languages. In terms of syntax, the pattern is very similar (Figure 2). However, the cross-linguistic semantic and syntactic similarity is rather moderate, meaning that English words which are difficult to translate into one language are not necessarily also difficult in other languages. In other words, there is partial overlap in terms of semantics and syntax between English and these six different target languages.

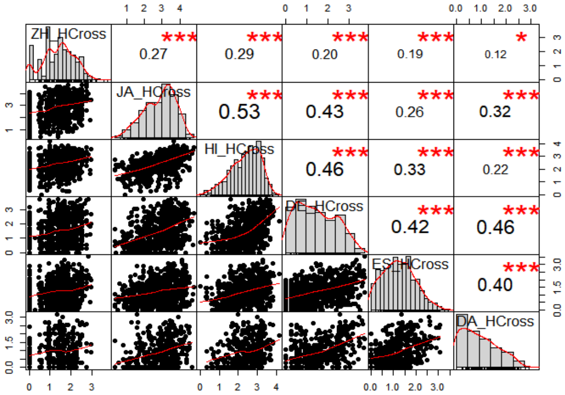


Figure 2. Syntactic cross-linguistic similarity as measured by the correlation of HCross values in 6 different target languages

In addition, we show that there are similarities and differences in how translators' behaviour in the different languages is affected by the semantic and syntactic overlap between English and the different target languages. Figure 3A shows that behaviour (typing duration per word) for Danish and Spanish translators is particularly affected by large word order differences in relation to the source (high *Cross* values), while Chinese remains mostly unaffected. In regard to semantic overlap between the source and the target languages (*HTra* Figure 3B), translators' behaviour (typing duration per word) working into Hindi is particularly affected while Danish, Japanese, Spanish and German are moderately affected, while Chinese is, again, only weakly affected.

The current study will explore these cross-linguistic similarities in the process and the product of translation in terms of the typological aspects specific to the language families involved. We will interpret the effect of word order and semantic differences across languages in translation in terms of models of the bilingual lexicon (de Groot 1992; Hartsuiker et al. 2004; Paradis 2004) and theories of lexical access (Collin & Loftus 1975; Poulisse & Bongaerts 1994) as well as with models of the human translation process (Halverson 2015).

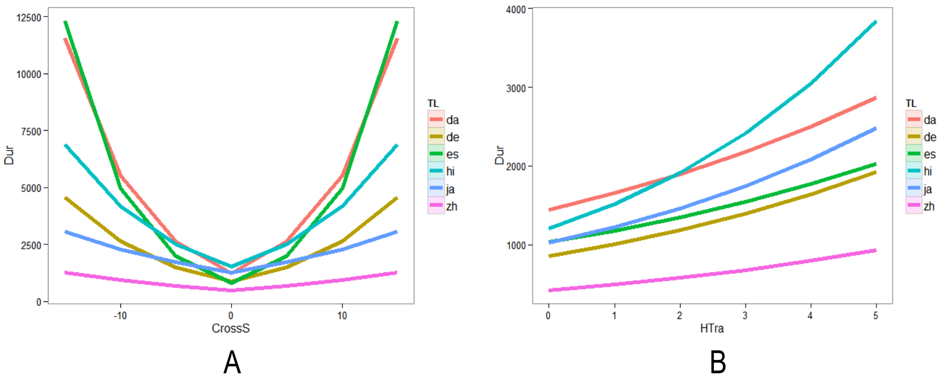


Figure 3. (A) The effect of Cross (word order differences) on production duration per word (in milliseconds) per target language. (B) The effect of HTra (predictability of target word) on production duration per word (in milliseconds) per target language.

References

Carl, M., Bangalore, S. & Schaeffer, M. J. (2016). *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. New York: Springer.

Carl, M. & Schaeffer, M. J. (2017). Why Translation Is Difficult : A Corpus-Based Study of Non-Literality in Post-Editing and From-Scratch Translation. *Hermes — Journal of Language and Communication Studies*, (56), 43-57.

Carl, M., Schaeffer, M. J. & Bangalore, S. (2016). The CRITT Translation Process Research Database. In M. Carl, S. Bangalore & M. Schaeffer (eds). *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. New Frontiers in Translation Studies. Cham, Heidelberg, New York, Dordrecht & London: Springer International Publishing, 13-54.

Collin, A. M. & Loftus, E. F. (1975). A Spreading Activation Theory of Semantic Processing. *Psychological Review* 82(6), 407-428.

de Groot, A. M. B. (1992). Determinants of Word Translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18(5), 1001-1018. doi:10.1037/0278-7393.18.5.1001.

Halverson, S. L. (2015). Cognitive Translation Studies and the merging of empirical paradigms. The case of 'literal translation'. *Translation Spaces* 4(2), 310-340.

Hartsuiker, R. J., Martin J. P. & E. Veltkamp. (2004). Is Syntax Separate or Shared between Languages? Cross-Linguistic Syntactic Priming in Spanish-English Bilinguals. *Psychological Science* 15(6), 409-414. doi:10.1111/j.0956-7976.2004.00693.x.

Malmkjær, K. (2011). Translation Universals. *The Oxford Handbook of Translation Studies*. Oxford: Oxford University Press, 83-93.

Paradis, M. (2004). *A Neurolinguistic Theory of Bilingualism*. Amsterdam & Philadelphia: John Benjamins.

Poulisse, N. & Bongaerts, T. (1994). First Language Use in Second Language Production. *Applied Linguistics* 15(1), 36-57.

Schaeffer, M.J. & Carl, M. (2013). Shared representations and the translation process: A recursive model. *Translation and Interpreting Studies*, 8(2), 169-190.

Schaeffer, M.J. & Carl, M. (2014). Measuring the Cognitive Effort of Literal Translation Processes. In U. Germann, M. Carl, P. Koehn, G. Sanchis-Trilles, F. Casacuberta, R. Hill & S. O'Brien (eds). *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT)*. Stroudsburg, PA: Association for Computational Linguistics, 29-37.

Schaeffer, M. J., Dragsted, B., Hvelplund, K. T., Winther Balling, L. & Carl, M. (2016). Word Translation Entropy: Evidence of Early Target Language Activation During Reading for Translation. In M. Carl, S. Bangalore & M. Schaeffer (eds). *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. Springer, 183-210. doi:10.1007/978-3-319-20358-4.

Tirkkonen-Condit, S. (2005). The Monitor Model Revised: Evidence from Process Research. *Meta* 50(2), 405-414.

Translationese, interpretese and foreignese: What do they have in common?

Marta Kajzer-Wietrzny
Adam Mickiewicz University, University of Bologna
mkajzer@wa.amu.edu.pl

The present paper will focus on the goals and the preliminary outcomes of a new project launched in March 2018 devoted to the study of translationese, interpretese and foreignese in the European Parliament in the Polish-English context.

The project involves a corpus-based and a corpus-driven analysis of lexical and grammatical features of native language, translationese and interpretese as well as foreignese, which will help to set apart the features potentially characteristic of the mentioned varieties. The study aims to verify the emerging hypotheses that the communicative situation of translation/interpreting and speaking a foreign language impose similar constraints on the language users, which may in turn affect the lexical and grammatical features of the linguistic output. It is assumed that the study helps to discover the characteristics of the language used in all the investigated communicative situations. This in turn will aid further research on the translation and interpreting process and translation and interpreting didactics. The analyses will be carried out on a corpus of debates in the European Parliament focusing on Polish-English translations and interpretations, as well as on the speeches delivered in English by the Polish native speakers. This source of data ensures high homogeneiety of the texts. In this particular dataset, all the source text consist in the speeches delivered at the European Parliament by the Polish MEPs. The speeches delivered in Polish are first interpreted simultaneously, which constitutes the input for the interpreting corpora; and then translated, which constitutes the input for the translation corpora. Speeches delivered by the Polish MEPs in English constitute the nonnative corpus in the investigated dataset.

The project stems from the tradition of descriptive translation studies dominated in recent decades by the ongoing search for translation universals (Baker 1993) which are supposed to reflect the specific cognitive setup of the translation process. The universals were originally approached from quite a radical perspective (Biel 2015) and it was believed that they were independent of the source language. Today, less restrictive approaches emerge, which consider the impact of e.g. interference (Mauranen 2004). Researchers have recently leaned towards the hypothesis that lexical and grammatical features earlier considered to be specific to translation may also appear in texts produced in other types of constrained communication (Chesterman 2004). Halverson (2003) suggests that such characteristics may be related to bilingual processing.

If indeed translation universals are related to bilingual processing, then similar distortions of lexical and grammatical features would be visible in the language of non-native language users. Studies on non-native and native texts point to phraseological differences (Gaspari 2013) and stylistic variations, e.g. revealed in the frequency and use of conjunctions (Gaspari & Bernardini 2008; Durham 2011; Wulf et al. 2014; Ivaska 2015). There are, however, few studies which next to non-native language take translation and interpreting into account in such analyses (Ferraresi et al. 2017), not to mention the Polish-English context.

Rabinovich et. al (2016) showed that when compared, corpora of translations and corpora of non-native texts are more similar to each other than to the corpus of native speeches. Such outcomes seem to confirm the hypotheses linking universals with constrained communication and bilingual processing.

Nevertheless, it is vital to expand the scope of analysis to interpreting, as it has been shown, that although similar to translation, in many respects interpreting is characterized by different lexical and grammatical tendencies (Shlesinger-Ordan 2012; Defrancq 2015; Bernardini et al. 2016; Ferraresi et al. 2017). Similarly, not only written but also spoken non-native texts should be included, as just like interpreting they are not subject to any editing.

The present project offers such a comprehensive research perspective. Its aim is to compare written and spoken corpora of translated, interpreted and non-native and native speeches delivered at the European Parliament, also to examine whether the lexical and grammatical features of interpreting into the foreign language and interpreting into the native tongue are similar or different.

The analysis will be both corpus-based and corpus-driven. A number of specific parameters will be treated as a starting point of the investigation. These refer both to lexical patterns (e.g. lexical density, list head coverage), grammatical patterns (e.g. frequency of articles or pronouns), stylistic features (the use of conjunctions) and phraseology (e.g. analysis of n-gramms and discourse organizers). The parameters have been selected based on the previous corpus-based studies on translation and nonnative language use (among other Laviosa 1998; Bernardini et al. 2016; Gaspari & Bernardini 2008; Defrancq et al. 2015; Rabinovich et al. 2016; Gaspari 2013; Durham 2011; Wulf et al. 2014; Pęzik 2011; Pęzik 2015). The examination will then enter the corpus-driven stage, whereby patterns emerging from the data will be analysed.

The texts and transcripts compiled in the course of the project will extend the language combination offered by the European Parliament Translation and Interpreting Corpus (EPTIC) developed at the University of Bologna (Bernardini et al. 2016; Ferraresi et al. 2017).

Research financed by the Polish Ministry of Science and Higher Education, Mobilność Plus (Mobility Plus) programme, grant number: 1610/MOB/V/2OI7/0.

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: John Benjamins, 233-250.

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

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

Biel, Ł. (2015). Translatoryka korpusowa. *Rocznik Przekładoznawczy* 10, 15-40.

Chesterman, A. (2004). Beyond the particular. In A. Mauranen & P. Kujamäki (eds). *Translation universals: Do they exist?* Amsterdam: John Benjamins, 33-49.

Defrancq, B., Plevoets, K. & Magnifico, C. (2015) Connective Items in Interpreting and Translation: Where Do They Come From? In J. Romero-Trillo (ed.) *Yearbook of Corpus Linguistics and Pragmatics 2015.* New York: Springer International Publishing, 195-222.

Durham, M. (2011). I think (that) something's missing: Complementizer deletion in nonnative e-mails. *Studies in Second Language Learning and Teaching* 1(3), 421-445.

Ferraresi, A., Bernardini, S. & Miličević, M. (2017). Words that go together. An exploration of the idiom principle in institutional spoken English. In *Corpus Linguistics* 2017, Birmingham, UK, 25-28 July 2017.

Gaspari, F. (2013). A phraseological comparison of international news agency reports published online: Lexical bundles in the English-language output of ANSA, Adnkronos, Reuters and UPI. *Varieng. Studies in Variation, Contacts and Change in English*, 13. http://www.helsinki.fi/varieng/series/volumes/13/gaspari/.

Gaspari, F. & Bernardini, S. (2008). Comparing non-native and translated language: Monolingual comparable corpora with a twist. In *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies*, Hangzhou, China, 215-234.

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

Ivaska, I. (2015). Tracing crosslinguistic influences in structural sequences: What does key structure analysis have to offer? *Bergen Language and Linguistics Studies* 6, 23-44.

Kajzer-Wietrzny, M. (2012). Interpreting universals and interpreting style. Unpublished PhD thesis. https://repozytorium.amu.edu.pl/jspui/bitstream/10593/
2425/1/Paca%20doktorska% 20Marty% 20Kajzer-Wietrzny.pdf (accessed 26 April 2017).

Kajzer-Wietrzny, M. (2018). Interpretese vs. Non-native Language Use: The Case of *Optional That*. In C. Bendazzoli, B. Defrancq & M. Russo (ed.) *Making way in Corpus-based Interpreting Studies. What do we know about interpreting thanks to corpora?* Berlin: Springer Verlag.

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

Lewandowska-Tomaszczyk, B. (2005). *Podstawy językoznawstwa korpusowego.* Łódź: Wydawnictwo UŁ.

Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (eds). *Translation universals: Do they exist?* Amsterdam: John Benjamins, 65-82.

Pęzik, P. (2011). Graph-based Analysis of Native and Learner Phraseology. *Proceedings of the 10th Teaching and Language Corpora Conference* (TALC10), Warsaw, Poland.

Pęzik, P. (2015). Using n-gram independence to identify discourse-functional lexical units in spoken learner corpus data. *International Journal of Learner Corpus Research* 1(2), 242-255.

Rabinovich, E., Nisioi, S., Ordan, N. & Wintner, S. (2016). On the Similarities Between Native, Non-native and Translated Texts. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016),* Berlin, Germany, 1870-1881.

Rabinovich, E. & Wintner, S. (2015). Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics (TACL)*, 3(1), 419-432.

Sandrelli, A. & Bendazzoli, C. (2005). Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference Series,* 1(1). https://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx.

Sergiu, N., Rabinovich, E., Dinu, L. P. & Wintner, S. (2016). A corpus of native, non-native and translated texts. *Proceedings of the Tenth International Conference on Language Resources and* Evaluation (LREC 2016), 23-28 May 2016, Portorož, Slovenia.

Shlesinger, M. & Ordan, N. (2012). More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target* 24 (1), 43-60.

Wulff, S., Lester, N. & Martinez-Garcia, M. T. (2014). That-variation in German and Spanish L2 English. *Language and Cognition* 6, 271-299.

Introducing the *International Comparable Corpus*

John Kirk[1], Anna Čermáková[2], Signe Oksefjell Ebeling[3], Jarle Ebeling[3], Michal Kren[4], Karin Aijmer[5], Vladimir Benko[6], Radovan Garabik[6], Rafal Gorski[7], Jarmo Jantunen[8], Marc Kupietz[9], Maria Simkova[6], Thomas Schmidt[9], Oliver Wicher[10]
University of Vienna[1], University of Birmingham[2], University of Oslo[3], Charles University[4],
University of Gothenburg[5], Slovak Academy of Sciences[6], Polish Academy of Sciences[7],
University of Jyvaskyla[8], Institute for the German Language[9], Paderborn University[10]
john.kirk@univie.ac.at, a.cermakova@bham.ac.uk, s.o.ebeling@ilos.uio.no, jarle.ebeling@usit.uio.no,
michal.kren@ff.cuni.cz, karin.aijmer@eng.gu.se, vladob@juls.savba.sk, garabik@kassiopeia.juls.savba.sk,
rafal.gorski@ijp.pan.pl, jarmo.h.jantunen@jyu.fi, kupietz@ids-mannheim.de, marias@korpus.juls.savba.sk,
thomas.schmidt@ids-mannheim.de, oliver.wicher@upb.de

This presentation introduces a new collaborative project: the *International Comparable Corpus* (ICC) (https://korpus.cz/icc), to be compiled from European national, standard(ised) languages, using the protocols for text categories and their quantities of texts in the *International Corpus of English* (ICE).

There is broad agreement that the ICE project has been highly successful because it has facilitated numerous comparisons of L1 and L2 national varieties of English worldwide. At the same time, spoken and/or written corpora have been compiled for other languages (cf. Xiao 2008). Corpus-based contrastive studies are a growing research area and researchers have voiced need for a more rigorous analytical framework, which includes adequate data sets. Contrastive analysis relies on data from mainly two types of corpora (Granger 2003): translation (parallel) corpora and comparable corpora (the terminology may differ, cf. McEnery & Xiao 2007). While parallel translation corpora contain original (source) texts and their aligned translations, comparable corpora contain texts that have been selected on comparable criteria (e.g. for text categories and often also quantities for each category). Comparable corpora may be monolingual or multilingual.

The ICC project is starting with nine European languages: Czech, English, Finnish, French, German, Norwegian, Polish, Slovak and Swedish. The underlying idea is that the different ICC national components will be mostly built up from existing corpus resources and the content of the individual components will be modelled on the ICE corpus, i.e. each ICC corpus will contain 1 million words as do the individual ICE corpora. The resources available (and their copyright status) for the individual languages are currently being reviewed. ICC will use TEI P5 XML as a common data format and it will also attempt to harmonize both markup and licensing of the individual national components.

The ultimate goal of the corpus project will be the facilitation of contrastive studies between English and other languages, but also between any other language combinations, involving highly comparable datasets of differing spoken and written registers. A striking feature of the ICC national components are their substantial spoken components (as in ICE corpora), at present comprising 600,000 words (or 60% of the total), with as many languages as possible providing transcriptions with audio-alignment. Such provision of spoken data across several comparable discourse situations for contrastive analysis will be unprecedented and invaluable for future research.

It has been agreed that the ICC corpus composition will be derived from the ICE design with the following adjustments. The ICC corpora will not contain in their spoken component two text categories comprising legal texts that are part of the ICE design. These categories are 'cross examinations' and 'legal presentations', which amount in the ICE corpora to 40,000 words. In the written component, the ICC corpora will not include text categories labelled in the ICE as 'non-printed'. However, there will be a new text category of blogs which will substitute the non-printed component amounting to 100,000 words. This category will be also newly compiled for English. The individual categories are represented by 2,000 word long extracts. The following table shows the individual text categories and their quantities which the ICC project team has agreed so far.

SPOKEN	**Words**	**WRITTEN**	**Words**
Dialogue/Conversation		**Printed**	
Direct conversation	180,000	Humanities (acad.)	20,000
Telephone conversation	20,000	Social sciences (ac.)	20,000
Class lessons	40,000	Natural sciences (ac.)	20,000
Broadcast discussions	40,000	Technical (acad.)	20,000
Broadcast interviews	20,000	Humanities (popular)	20,000
Parliament debates	20,000	Social sciences (pop.)	20,000
Business transactions	20,000	Natural scienc. (pop.)	20,000
Monologue		Technical (popular)	20,000
Spontaneous commentaries	40,000	Reportage	40,000
Unscripted speeches	60,000	Instruct. (admin.)	20,000
Demonstrations	20,000	Instruct. (hobbies)	20,000
Broadcast news	40,000	Press editorials	20,000
Broadcast talks	40,000	Fiction	40,000
Speeches (not broadcast)	20,000	**Web**	
		Blogs	100,000

The ICC team is currently represented by the following institutions. For the ICE corpus by University of Vienna, for Czech by the Insitute of the Czech National Corpus at Charles University, for Finnish by University of Jyvaskyla, for French by Paderborn University, which is one of the institutions participating in the collection of the *Corpus de référence du français contemporain* (CRFC), for German by the Institute for the German Language in Mannheim, for Norwegian by University of Oslo, for Polish by the Polish Academy of Sciences, for Slovak by the Slovak Academy of Sciences where the *Slovak National Corpus* is based, and for Swedish by University of Gothenburg. The individual languages have currently access to various amounts of resources, in many cases most of the resources are already available but in some cases they will need to be collected. Additional information can be found at the project website: https://korpus.cz/icc.

References

Granger, S. (2003). The Corpus Approach: A common way forward for contrastive linguistics and translation studies? In S. Granger, J. Lerot & S. Petch-Tyson (eds). *Corpus-based Approaches to Contrastive Linguistics*. Amsterdam: Rodopi, 17-29.

McEnery, T. & Xiao, R. (2007). Parallel and Comparable Corpora: What is Happening? In G. M. Anderman & M. Rogers (eds). *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters, 18-31.

Xiao, R. (2008). Existing Corpora. In A. Lüdeling & M. Kytö (eds). *Corpus Linguistics: An International Handbook*. Berlin: Walter de Gruyter, 383-456.

Using comparable corpora for translating complex noun groups in specialised texts (from English to French)

Natalie Kübler, Alexandra Mestivier, Mojca Pecman
CLILLAC – ARP, Université Paris Diderot
nkubler@eila.univ-paris-diderot.fr, avolansk@eila.univ-paris-diderot.fr, mpecman@eila.univ-paris-diderot.fr

The present study focuses on one type of difficulty that translation students encounter when working with specialised texts, namely the comprehension and translation of heavy, complex noun phrases (NP). The study also presents the corpus-based methodology for translation training which we have implemented relying on Zanettin (1998), Aston (1999), Maia (2003), and many others. This methodology was devised more that a decade ago at the Translation Department of Paris Diderot University, and has systematically been evaluated and improved since 2013. The overall framework for teaching specialised translation (ST) with corpora involves a vast number of competences and disciplines which are integrated in the curriculum, such as corpus linguistics (corpus compilation, annotation and querying), terminology management, collaboration with domain experts, literature and information retrieval, and ST, among others. To put this framework into practice we created an experimental protocol in collaboration with the Earth and Planetary Sciences (EPS) department at the University Paris Diderot. Within this protocol, students translate fragments of scientific articles in EPS, published in high-impact, peer-reviewed journals such as *Nature*, *Science*, *Earth and Planetary Science Letters*, following the in-depth study of key terms used in the articles and their terminographical processing in an in-house term-base, ARTES.

In 2013 we began evaluating the impact of corpus use on the quality of student translations. We wanted to identify the most frequent translation errors made in this type of ST and whether these errors could be avoided by using corpora. We therefore further refined our methodology to include an evaluative phase (Figure 1): every year the students perform two translation tasks, the first one using online dictionaries exclusively, and the second one with the additional help of EPS comparable corpora. The produced translations are uploaded on a BRAT server[1] and annotated using the MeLLANGE translation error typology[2] (Secara 2005; Castagnoli et al. 2011). The two translation collections are then compiled into two learner translation sub-corpora (SP-TRANS1 and SP-TRANS2) according to the two conditions of production.

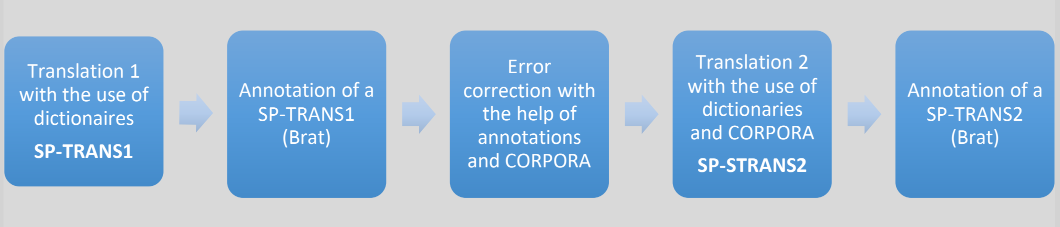

Translation 1 with the use of dictionaires
SP-TRANS1
Annotation of a SP-TRANS1 (Brat)
Error correction with the help of annotations and CORPORA
Translation 2 with the use of dictionaries and CORPORA
SP-STRANS2
Annotation of a SP-TRANS2 (Brat)

Figure 1. Evaluation steps of our corpus-based specialised translation training methodology

Between SP-TRANS1 and SP-TRANS2 tasks we annotate translations gathered in SP-TRANS1 and provide input to show (a) how corpus querying may help to identify equivalent terms and identify means of couching terms using a phraseology that is appropriate for the target language, and (b) how corpus use can help to avoid different types of translation errors. Before the second task, students produce a second version of the first translation on the basis of the error annotation, explaining for each modified segment how a better translation solution was found (providing corpus queries and concordance lists on which their choices were based). This intermediate step is meant to enhance their awareness of corpus efficiency for translation.

[1] http://brat.nlplab.org.
[2] http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg.

The most salient and frequent types of errors identified through quantitative analysis represent useful material for investigating ST difficulties. In Kübler et al. (2015) we have quantified the improvements that students achieve by using the EPS corpora in their translation task by comparing normalised (PMW) frequencies of each error-type in SP-TRAN1 and SP-TRAN2. In Kübler et al. (2016), we adopted a textometric approach and measured the specificities of each type of error in the two sub-corpora. In both cases we found that some error-types (such as *wrong term equivalent*, *wrong collocation* or *wrong preposition*) are better candidates for improvement through corpora use than other error-types (such as *literal translation*, *syntax error* or *distortion*).

Aiming to find out how to exploit corpora more efficiently for the error-types which seem corpus-insensitive, in the present contribution we have a closer look at one particular error-type for which little improvement is obtained in the SP-TRANS2, namely *distortion*. Among the many configurations in which distortion occurs, we identified a recurring pattern: the occurrence of heavy, complex NPs, for which students fail to identify the head (Table1).

Source text	Student translation
*[…] the presence of atmospheric noble gases in **subduction-zone serpentinites** […]*	**[…] la présence des gaz nobles atmosphériques dans les **zones de subduction constituées de serpentinite** […]*
*[…] acquiring noble gas concentration data for **hydrous subduction zone minerals** […]*	**[…] collecter des données de concentration de gaz nobles dans **les zones de subduction de minéraux hydratés** […]*

Table 1. Examples of heavy, complex NPs that represent a difficulty in ST

During the academic year 2017-2018, we have devised a classroom activity to increase student awareness on this issue and help them use corpora to a) confirm the NP interpretation, i.e. identify the head and provide a syntactic representation of the complex NP, and b) suggest the most appropriate way(s) of translating the NP. We have also added a finer-grained error category, *complex NP*, to the MeLLANGE typology in order to allow us to single out cases where students failed to identify the NP head and hence mistranslated the NP.

The present contribution will thus focus on the cases of distortion due to the erroneous syntactic interpretation of complex NP structure, and on the quantitative evaluation of the impact that the pedagogical input based on corpus use can produce on this specific error-type. This evaluation will take place during the February-April 2018 period. Consequently, our study will provide a relevant contribution to two research topics: corpus use in translation teaching and nominal compounding in LSPs. A number of recent works have reported on the usefulness of corpora for translation training (Bowker & Bennison 2003; Castagnoli et al. 2011; Loock et al. 2014; Frankenberg-Garcia 2015) but this field of research is only just expanding, and we need further evidence on methods and procedures for efficient corpora integration in ST classes. In quite the same way, the tendencies of ESP for heavy nominal compounding and its consequences for translation have been studied by few linguists (Maniez 2008, 2010, 2013, 2017; Portelance 1987) and deserve larger attention, all the more so since Mestivier's study (2015) has demonstrated the increase of heavy compounding in specialised texts over last decades.

References

Aston, G. (1999). Corpus use and learning to translate. *Textus* 12, 289-313.

Bowker, L. & Bennison, P. (2003). Student Translation Archive and Student Translation Tracking System. Design, Development and Application. In F. Zanettin, S. Bernardini & D. Stewart (eds). *Corpora in translator* education. Manchester: St. Jerome Publishing, 103-118.

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.

Frankenberg-Garcia, A. (2015). Training translators to use corpora hands-on: Challenges and reactions by a group of 13 students at a UK university. *Corpora* 10(3), 351-380.

Kübler, N., Mestivier, A., Pecman, M. & Zimina, M. (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*, 7-10 June 2016, Nice, France. [http://jadt2016.sciencesconf.org/82617/document].

Kübler, N., Pecman, M., Volanschi, A. M. (2015). Étude sur l'utilisation des corpus dans l'enseignement de la terminologie et de la traduction spécialisée. *Terrains de recherche en linguistique appliquée* (TRELA 2015), July 2015, Paris, France. [trela.clillac-arp.univ-paris-diderot.fr].

Loock, R., Mariaule, M. & Oster, C. (2013). Traductologie de corpus et qualité : Étude de cas. *Tralogy II*, Session 5 — Assessing Quality in MT / Mesure de la qualité en TA, 17-18 January 2013, Paris, France. [http://lodel.irevues.inist.fr/tralogy/index.php?id=243].

Maia, B. (2003). Training translators in terminology and information retrieval using comparable and parallel corpora. In F. Zanettin, S. Bernardini & D. Stewart (eds). *Corpora in Translator Education*. Manchester: St. Jerome, 43-54.

Maniez, F. (2008). Using the Web and corpora as language resources for the translation of complex noun phrases in medical research articles. *Panacea* IX(26), 162-167. [http://medtrad.org/panacea/IndiceGeneral/n26_tribunaManiez.pdf].

Maniez, F. (2010). La traduction des adjectifs composés en langue médicale : Étude d'un corpus bilingue anglais-français. *Séminaire du CRTT*, Université de Lyon 2, Lyon.

Maniez, F. (2013). The translation into French of adjectives formed with a noun and a past participle in English-language medical articles. *Panacea* XIV(38), 240-247. [http://www.tremedica.org/panacea/IndiceGeneral/n38tradyterm_ManiezF.pdf].

Maniez, F. (2017). Évaluation des récentes avancées de la traduction automatique : le cas des adjectifs composés formés à partir d'un participe passé en anglais de spécialité. *ASp* 72, 29-48

Mestivier, A. (2015). Productivity and Diachronic Evolution of Adjectival and Participial Compound Pre-modifiers in English for Specific Purposes. *Fachsprache* XXXVII(1-2), 2-23.

Portelance, C. (1987). Fertilisation terminologique ou insémination terminologique artificielle ? *Meta* 32(3), 356-360.

Secară, A. (2005). Translation Evaluation – A State of the Art Survey. *Proceeding of the* eCoLoRe*/MeLLANGE Workshop*, 21-23 March 2005, Leeds, England. Translation Studies Abstracts. St. Jerome Publishing, 39-44.

Zanettin, F. (1998). Bilingual Comparable Corpora and the Training of Translators. *Meta* 43(4), 616-630. doi:10.7202/004638ar.

Recent developments in the European Reference Corpus (EuReCo)

Marc Kupietz[1], Ruxandra Cosma[2], Dan Cristea[3,4], Nils Diewald[1],
Beata Trawiński[1], Dan Tufiş[5], Tamás Váradi[6], Angelika Wöllstein[1]
Institut für Deutsche Sprache[1], University of Bucharest[2],
Romanian Academy Iaşi[3], "Alexandru Ioan Cuza" University of Iaşi[4],
Institute for Artificial Intelligence Mihai Drăgănescu[5], Hungarian Academy of Sciences[6]
kupietz@ids-mannheim.de, ruxandra.cosma@lls.unibuc.ro, dcristea@info.uaic.ro,
diewald@ids-mannheim.de, trawinski@ids-mannheim.de, tufis@racai.ro,
varadi.tamas@nytud.mta.hu, woellstein@ids-mannheim.de

Introduction

The past 20 years have seen an emergence of national, reference and other large corpora of numerous European languages (cf. Kupietz et al. 2017). Most of them have been or are being built in projects of limited duration, but typically based at institutions that are at least to some degree responsible for curating data and for making it available to the respective scientific communities also after the building phase. The idea of EuReCo (Kupietz et al. 2017) is that such institutions should join forces to develop techniques that allow for a unified view on the existing corpora and to use them as a base for comparable corpora. The common infrastructure will include the following main features: metadata shall use attributes and values that are mappable among all pairs of component corpora, annotation conventions shall be harmonised (as much as linguistic idiosyncrasies permit) to the point that comparable queries on different languages shall be possible and shall produce comparable results, textual content of the component corpora can remain at their hosting institution and be locally extended (quantitatively), updated (w.r.t textual data) or upgraded (w.r.t. metadata and annotation), any component can be accessed from anywhere through a Portal entry point, mixed screens combining more comparable searched for material can be activated dynamically, tools doing statistical counts can be invoked by users on any language component and on any combinations of them, tools used in statistical counts and comparisons shall be able to combine flows of data contributed by all local linguistic data hosts, etc. The expected advantages of this approach are that no comparable corpora would have to be built from scratch, all existing corpora can remain at their hosting institutions – avoiding IPR and licensing issues – and the base for the selection of comparable pairs of sub-corpora could directly benefit from the expansion of the individual initial corpora. The downside of this approach, however, compared e.g. to the similar approach of the ICC (Kirk & Čermáková 2017) is of course that the stratification and composition of possible comparable corpus pairs cannot be designed in advance, but rather depends on the strata manifested by metadata in the source corpora, their respective sizes and the translatability of these stratifications or rather metadata taxonomies between individual corpus pairs.

Previous and current work

EuReCo is currently based on the following corpora:

- The German Reference Corpus DeReKo (Deutsches Referenzkorpus), with more than 42 billion words (Kupietz & Lüngen 2014; Kupietz et al. to appear), the largest linguistically motivated collection of German texts, featuring a so-called primordial-sample design, which is also fundamental for the definition of different virtual comparable corpora in the EuReCo context.
- The Reference Corpus of Contemporary Romanian Language CoRoLa, containing almost one billion words, which was publicly launched in December 2017 and can be queried via different interfaces, including KorAP.[1]
- The Hungarian National Corpus HNC, that has recently been substantially upgraded and extended to gigaword size (Váradi 2002; Oravecz et al. 2014).

KorAP

[1] http://corola.racai.ro/

The current technical basis for EuReCo is the corpus query and analysis platform KorAP[2] that has recently been developed at the IDS (Bański et al. 2013; Diewald et al. 2016). KorAP is the designated successor of the corpus search and management system COSMAS,[3] which is currently used by 40,000 researchers working on the German language. The features of KorAP that are essential for EuReCo are particularly (i) its ability to manage corpora that are physically located at different places, in order to comply with typical license restrictions (cf. Kupietz et al. 2014) and (ii) its ability to dynamically create virtual sub-corpora based on text properties and to manage these virtual corpora in a persistent way, for example allow for reusability and reproducibility.

DRuKoLA: The first EuReCo blueprint

Parts of the EuReCo vision have already been implemented in the DRuKoLA-project,[4] which is centered around DeReKo and CoRoLa (Cosma et al. 2016). One of its main objectives is to provide a common platform for constructing various kinds of comparable corpora based on text properties and to analyse them for contrastive linguistic purposes.

The present state of the art of DRuKoLA relevant to EuReCo is that CoRoLa can be accessed publicly via KorAP and that a first virtual comparable corpus is defined. This first definition is based solely on a mapping from CoRoLa's two-level topic domain taxonomy to DeReKo's topic domain taxonomy (also two-levelled, see Klosa et al. 2012: 88). In order to be able to map a sufficiently substantial portion from the smaller corpus CoRoLa, it was necessary to map from top-level domains to sub-level domains and vice versa.

DeutUng

As a second EuReCo pilot project, DeutUng[5] has recently started to integrate the Hungarian National Corpus (HNC) into EuReCo. With respect to the establishment of an infrastructure and research methodology for comparable corpora, DeutUng is similar to DRuKoLA.[6]

Relevance for Cross-Linguistic Research

Cross-linguistic research needs multilingual data. So far, parallel / translational resources have played a major part, both in contrastive linguistics and language typology as well as in translational studies and foreign language education (cf. James 1980; Chesterman 1998; Granger et al. 2003; Granger 2010; Johansson 2007; Cysouw & Wälchli 2007). While the usefulness of parallel resources for cross-linguistic research is obvious (as they provide data that convey the same meaning and can thus serve as a basis for establishing equivalence between entities across different languages), they show a number of undesirable effects. The problems include particularly the so-called *source language shining through* (Teich 2003), and other specifics of translated texts, such as *over-normalization, simplification,* etc. No such problems arise in comparable data. In this respect, EuReCo, which is based on the existing national or reference corpora, provides a unique linguistic resource offering new perspectives for fine grained contrastive research on authentic cross-linguistic data, applications in translation studies and foreign language teaching and learning.

References

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C. & Witt, A. (2013). KorAP: the new corpus analysis platform at IDS Mannheim. In Z. Vetulani, H. Uszkoreit & M. Kubis (eds). *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference* (LTC 2013). Poznań, Poland: Fundacja Uniwersytetu im. A., 586-587.

Chesterman, A. (1998). *Contrastive Functional Analysis*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

[2] https://korap.ids-mannheim.de/

[3] http://cosmas2.ids-mannheim.de

[4] DRuKoLA (2016-2019) is funded by the Alexander von Humboldt-Foundation, as a Research Group Linkage Programme. The acronym combines central goals of the project: corpus development and contrastive linguistic analysis (*Sprachvergleich korpustechnologisch. Deutsch-Rumänisch*).

[5] DeutUng (2017-2020) is a cooperation project between IDS Mannheim and the University of Szeged with the Research Institute for Linguistics at the Hungarian Academy of Sciences as associated partner. It is also funded by the Alexander von Humboldt-Foundation as a Research Group Linkage Programme.

[6] With respect to linguistic application, however, DeutUng has as an additional focus on second language acquisition.

Cosma, R., Cristea, D., Kupietz, M., Tufiş, D. & Witt, A. (2016). DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora. In P. Bański, M. Kupietz, H. Lüngen, A. Witt, A. Barbaresi, H. Biber, E. Breiteneder & S. Clematide (eds). *4th Workshop on Challenges in the Management of Large Corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), May 23-28, Portorož, Solevia. Paris: European Language Resources Association (ELRA), 28-32.

Cysouw M. & Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung* 60(2), 95-99.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P. & Witt, A. (2016). KorAP Architecture – Diving in the Deep Sea of Corpus Data. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds). *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), May 23-28, Portorož, Solevia. Paris: European Language Resources Association (ELRA), 3586-3591.

Granger, S. (2010). Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Journal of Shanghai Jiaotong University*, 2, 14-21.

Granger, S., Lerot, J. & Petch-Tyson, S. (eds). (2003). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Amsterdam & Atlanta: Rodopi.

James, C. (1980). *Contrastive Analysis*. London: Longman.

Johansson, S. (2007). *Seeing through multilingual corpora. On the use of corpora in contrastive studies*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Kirk, J. & Čermáková, A. (2017). From ICE to ICC: The new International Comparable Corpus. In P. Bański, M. Kupietz, H. Lüngen, P. Rayson, H. Biber, E. Breiteneder, S. Clematide, J. Mariani, M. Stevenson & T. Sick (eds). *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing*. Mannheim: Institut für Deutsche Sprache, 7-12.

Klosa, A., Kupietz, M. & Lüngen, H. (2012). Zum Nutzen von Korpusauszeichnungen für die Lexikographie. *Lexicographica* 28, 71-97.

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds). *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC 2010), May 17-23, Valletta, Malta, 1848-1854.

Kupietz, M., Lüngen, H., Bański, P. and Belica, C. (2014). Maximizing the Potential of Very Large Corpora. In M. Kupietz, H. Biber, H. Lüngen, P. Bański, E. Breiteneder, K. Mörth, A. Witt & J. Takhsha (eds). *Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora* (CMLC2). Paris: European Language Resources Association (ELRA), 1-6.

Kupietz, M., Lüngen, H., Kamocki, P. and Witt, A. (to appear in 2018). The German Reference Corpus DeReKo: New developments – new opportunities. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), 7-12 May 2018, Miyazaki, Japan.

Kupietz, M., Witt, A., Bański, P., Tufiş, D., Cristea, D. & Váradi, T. (2017). EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In P. Bański, M. Kupietz, H. Lüngen, P. Rayson, H. Biber, E. Breiteneder, S. Clematide, J. Mariani, M. Stevenson & T. Sick (eds). *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing*. Mannheim: Institut für Deutsche Sprache, 15-19.

Oravecz, Cs., Váradi, T. & Sass, B. (2014). The Hungarian Gigaword Corpus. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds). *Proceedings on the Ninth International Conference in Language Resources and Evaluation* (LREC 2014), May 26-31, Reykjavik, Iceland. Paris: European Language Resources Association (ELRA), 1719-1723.

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

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ș. D. & Boroş, T. (2016). The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds). *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), May 23-28, Portorož, Solevia. Paris: European Language Resources Association (ELRA), 2516-2521.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ș. D., Boroş, T., Teodorescu, N. H., Cristea, D., Scutelnicu, A., Bolea, C., Moruz, A. & Pistol, L. (2015). CoRoLa Starts Blooming – An Update on the Reference Corpus of Contemporary Romanian Language. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen & A. Witt (eds). *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora* (CMLC-3). Mannheim: Institut für Deutsche Sprache, 5-10.

Váradi, T. (2002). The Hungarian National Corpus. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas & Paris: European Language Resources Association (ELRA): 385-389.

Translation universals: Evidence from a study of Croatian and Italian translated texts

Ivana Lalli Paćelat
Juraj Dobrila University of Pula
ilalli@unipu.hr

Corpus-based translation studies have shown that translated texts differ from non-translated texts and that, independently of the language, translations show some common features (e.g. Baker 1996; Bernardini 2011; Laviosa 2002; Xiao 2010). It is still largely debated whether they should be considered as translation universals or just general tendencies in translated texts (cf. Bernardini & Zanettin 2004; Chesterman 2004; Mauranen 2008; Teich 2003; Xiao 2010; Xiao & Dai 2014). Research has also been conducted on the differences between registers in translated and non-translated texts and across languages suggesting different methodological approaches (e.g. Biber 1995; Neumann 2010; Teich 2003). An insufficient number of studies on translation universals, which does not include the English language (Xiao 2010; Jiang & Tao 2017), and particularly those involving Croatian and Italian-Croatian language pair, was the motivation for this study.

The research presented here was carried out in the framework of my PhD project concerned with the investigation of the nature of the legislative register in translated and original texts in the language pair Croatian-Italian. This paper explores the existence of general tendencies in translated texts or of translation universals and their nature. The hypothesis predicts that, given the existence of universal translation features, the translated texts are more similar to one another than parallel texts of related languages. Furthermore, the research aims at finding out whether the translations have the same lexico-grammatical features as the target language of the studied register or they belong to a special register.

According to Biber (1995) the basic requirements for the register analysis are (i) the comparative approach, (ii) the quantitative analysis and (iii) a representative sample. In order for these requirements to be met, six corpora belonging to four different corpus types were employed for the study; firstly, reference corpora for both languages: (1) Croatian National Corpus (HNK v 3.0) and (2) Corpus di Italiano Scritto (CORIS); secondly, (3) specialized bilingual comparable corpus composed of national legislative documents in both languages (subcorpora of HNK v 3.0 and CORIS); thirdly, (4, 5) monolingual corpora of original national legislative documents and translations of legislative documents of the European Union in the same language used as comparable corpus and lastly, a (6) parallel corpus consisting of Croatian and Italian translations of legal documents of the European Union. For the description of corpus parameters for HNK see Tadić (2009) and for CORIS Rossini Favretti et al. (2002). The approach adopted in this study was a hybrid one, without an 'a priori' established theoretical framework, but the corpora were annotated at part of speech (PoS) and lemma level. The selection of the linguistic features for the quantitative analysis followed previous studies (e.g. Cortelazzo 2013; Neumann 2010; Teich 2003; Venturi 2011; Xiao & Dai 2014), and was driven by primary corpus obtained data. In order to investigate the properties of translated texts, considered as a special register type, linguistic features at both lexical and grammatical level were quantitatively analysed and statistically evaluated among all the corpora and the two languages in question. Although the six corpora included in the research were comparable with respect to size, purpose and structure, it was indispensable, due to the nature of the planned quantitative analysis, to make them comparable at the POS and morphosyntactic description (MSD) tagging level. Hence, several analyses and procedures were needed, including a detailed contrastive analysis of the two languages, in order to achieve comparability, greater reliability and accuracy of results (Lalli Paćelat 2016). The analysis was performed by using NoSketch Engine (Rychlý 2007) for HNK v3.0 (Tadić 2009) and for CORIS the on-line interface designed by F. Tamburini. Each result was interpreted and assigned to the corresponding translation universal.

The Italian translational corpus showed the tendency towards normalization and the Croatian translational corpus towards leveling out. Both these translation universals were in this study understood somewhat differently in contrast to the study carried out by Baker (1996). Leveling out is usually defined as the tendency of the translation, according to some ratios and values, to be equidistant from two extreme poles. While in previous research these two extremes were represented by the source language and the target language or the source texts and the target texts, in this study they are represented by the neutral register and the legislative register of the target language. Namely, the values of the translational corpora gravitate between the values of the general reference corpora and the values of the specialized corpora. Normalization usually refers to the translator's tendency towards excessive use of typical patterns of the target language. In this study it was not about typical patterns of the target language in general, but of the target register since the values of translational corpora are largely moving away from the values of the reference corpora. The normalization can therefore be attributed to more frequent use of those lexical and grammatical features that are considered typical of the legislative register of both languages.

The Italian and the Croatian translational corpus showed the same tendency only in 5 of 29 observed features. This confirms that two translations from the same source language/texts do not behave in the same way and that the quantity and the type of the translation universals depend on the morphosyntax of the target language and that translations are less susceptible to the influence of the source language. The research showed and confirmed that translation universals are neither easy to identify nor unambiguous to interpret (Xiao & Dai 2014). Universal translation features was found in both languages, but not always the same features and not with the same frequency. However, these features do not make the translated texts considerably different from comparable non-translated texts in the same language. The largest number of similarities found between specialized and translational corpora in the same language confirms the authenticity of the translations and their orientation towards the target language, and in particular, towards the features of the target register.

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, 175-187.

Bernardini, S. (2011). Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS* 26, 2-13.

Bernardini, S. & Zanettin, F. (2004). When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In A. Mauranen & P. Kuyamaki (eds). *Translation Universals: Do they Exist?* Amsterdam: John Benjamins, 51-62.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Chesterman, A. (2004). Beyond the particular. In A. Mauranen & P. Kujamäki (eds). *Translation Universals: Do They Exist?* Amsterdam: John Benjamins, 33-49.

Cortelazzo, M. A. (2013). Leggi italiane e direttive europee a confronto. In *Realizzazioni testuali ibride in contesto europeo. Lingue dell'UE e lingue nazionali a confronto*. Trieste: EUT – Edizioni Università di Trieste, 57-66.

Jiang, Z. & Tao, Y. (2017). Translation Universals of Discourse Markers in Russian-to-Chinese Academic Texts: A Corpus-based Approach. *Zeitschrift für Slawistik* 62(4), 583-605.

Lalli Paćelat, I. (2016). Priprema usporedivih korpusa za usporedbu. In T. Erjavec & D. Fišer (eds). *Proceedings of the Conference on Language Technologies & Digital Humanities*. Ljubljana: Ljubljana University Press, 111-120.

Laviosa, S. (2002). *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam & Atlanta: Rodopi.

Mauranen, A. (2008). Universal tendencies in translation. In G. Anderman & M. Rogers (eds). *Incorporating Corpora. The Linguist and the Translator*. Clevedon: Multilingual Matters, 32-48.

Neumann, S. (2010). Quantitative Register Analysis Across Languages. In E. Swain (ed.) *Thresholds and Potentialities of Systemic Functional Linguistics: Multilingual, Multimodal and Other Specialised Discourses*. Trieste: EUT Edizioni Università di Trieste, 85-113.

Rossini Favretti, R., Tamburini, F.& De Santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model". In A. Wilson, P. Rayson & T. McEnery (eds). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Munich: Lincom-Europa, 27-38.

Rychlý, P. (2007). A Modular Corpus Manager. In P. Sojka & A. Horák (eds). *First Workshop on Recent Advances in Slavonic Natural Language Processing* (RASLAN 2007). Brno: Masaryk University, 65-70.

Tadić, M. (2009). New version of the Croatian National Corpus. In D. Hlaváčková, A. Horák, K. Osolsobě & P. Rychlý (eds). *After Half a Century of Slavonic Natural Language Processing*. Brno: Masaryk University, 199-205.

Teich, E. (2003). *Cross-linguistic variation in system and text*. Berlin & New York: Mouton de Gruyter.

Venturi, G. (2011). *Lingua e diritto: una prospettiva linguistico-computazionale*. Unpublished PhD thesis. University of Turin. Available online at: http://www.italianlp.it/?page_id=81 [15.09. 2013].

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

Xiao, R. & 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.

Dictionary, corpus and CAT tool use in legal translation: A comparative pilot study of student translations

An Lambrechts, Heidi Verplaetse
Katholieke Universiteit Leuven
an.lambrechts@kuleuven.be, heidi.verplaetse@kuleuven.be

Theoretical background

In an early study conducted by Bowker (1998) it was shown that the use of monolingual original corpora in the target language, viz. corpora containing texts in one language produced by native speakers, leads to an increase in translation quality, including understanding of the subject field, selecting correct terminology and idiomatic expressions (Bowker 1998: 648). This may be due to the fact that these monolingual original corpora do not contain translated language, and are therefore uninfluenced by so-called translationese. However, since Bowker (1998) the impact of monolingual original corpora on translation quality has not been explored further to any great extent.

In addition, survey research has indicated a low level of corpus awareness among student and professional translators (Zaretskaya et al. 2015: 250), which can be attributed to the lack of training among future translators for the use of corpus tools (e.g. WordSmith, AntConc, Sketch Engine), the unavailability of ready-made specialized corpora (Wilkinson 2010) or unawareness among translators about the potential benefits of corpora “as a supplement to other resources and references” (Frankenberg-Garcia 2015: 354). The use of computer-assisted translation (CAT) tools is currently much more widespread than corpus use in translation practice. Therefore, user-friendly integration of (monolingual original) corpora into CAT tools, particularly in their translation memory (TM) component, might allow and encourage (future) translators to use, construct and query such target language corpora better as part of their workflow (cf. Bernardini & Castagnoli 2008). Furthermore, these corpora can complement the CAT tool translation approach, as TMs also entail disadvantages: TM segments, like most conventional dictionary entries, are decontextualized and Jiménez-Crespo (2009) has shown that inconsistency is higher in translations executed with TMs than in original texts. As translation quality depends on the content of the TM, TMs of poor quality will also negatively impact overall translation quality. To address these drawbacks original language corpora may be incorporated in translation practice.

Aims and methods

In order to assess whether the use of monolingual original corpora in the target language proves more beneficial with regard to translation quality than the use of conventional bilingual resources (bilingual dictionaries) and CAT tools (TMs), a pilot study was conducted among 11 master students taking a specialized legal translation course. We chose specialized texts as these are characterized by subject-specific language or Language for Specific Purposes (LSP). The use of corpora benefits translation of such texts greatly (Kübler 2003: 25). The students were asked to translate two legal texts fragments. Their translation activity was also logged using keylogging software (Inputlog) and screen recording software (ActivePresenter). Text fragment one was translated with either a bilingual dictionary or a specialized TM in SDL Trados Studio 2017. Subsequently, a corpus compilation phase took place. During this phase, which lasted 30 minutes at most, the students were asked to compile a monolingual corpus of texts originally produced in the target language using pre-defined keywords. In order to find resources which were originally written in the target language, website country codes could be adjusted. For the translation of the second text fragment, the students either used a bilingual dictionary in combination with their self-compiled monolingual corpus in the target language or a specialized TM and their self-compiled corpus. The corpus was queried using the software AntConc. Translation quality assessment was initially error-based and conducted using error typologies based on MeLLANGE (Kübler et al. 2016) and on a set of annotation guidelines defining different error categories (Daems & Macken 2013). In a later stage the translations will also be assessed from a linguistic perspective using tagging and parsing procedures.

Preliminary pilot test results

As revealed by the data from the screen recording software ActivePresenter, the number of corpus files gathered varied greatly, but the more corpus files the students had at their disposal in their self-compiled corpus, the more they consulted it. This led to longer translation times, but also enabled these students to maintain the same quality translation standard they achieved without corpus resources (text fragment 1). An increased number of corpus consultations also appeared to correlate with lower error rates in comparison to students with fewer corpus files and fewer corpus consultations.

No major difference could be seen in the number of errors made between TM only translations and TM-corpus translations: the number of errors made was only slightly higher. Looking at error frequencies it was established that most frequently words and phrases were translated correctly in TM only and dictionary-corpus translations, but the translations did not fit the context (word sense disambiguation). This implies that nor a (decontextualized) TM, nor a (contextualized) corpus, when used separately, provide enough context to determine a correct context-specific translation. However, in TM-corpus translations the error category word sense disambiguation occurred much less often than in TM only translations: words and phrases were translated correctly more often and the translations used fitted the context. This suggests that the combination TM-corpus accounts for an increased number of adequate translations: by consulting the TM/corpus, the context-specific translation is found. However, the error category mistranslation (viz. when a word or phrase is not a possible translation of that word or phrase) also ranks high under the different translation conditions (including TM-corpus translations): the translation is unknown, nor can it be retrieved from the aids allowed depending on the translation condition. On the one hand, the above suggests that the combination TM-corpus may positively influence translation quality with regard to error frequencies (word sense disambiguation occurring less frequently). On the other hand, there are still other factors at play which may increase the error rate (e.g. the number of mistranslations), such as limitations of the TM/corpus, and inadequate corpus compilation and searching skills.

The results to be presented for this paper will also include test results from a replication of the experiment described above with other student groups, including 11 master students taking a specialized scientific/medical translation course.

References

Bernardini, S. & Castagnoli, S. (2008). Corpora for translator education and translation practice. In E. Yuste Rodrigo (ed.) *Topics in Language Resources for Translation and Localisation*. Amsterdam & Philadelphia: John Benjamins, 39-65.

Bowker, L. (1998). Using specialized monolingual native-language corpora as a translation resource: A pilot study. *Meta: Journal des traducteurs / Meta: Translators' Journal*, 43(4), 631-651.

Daems, J. & Macken, L. (2013). Annotation Guidelines for English-Dutch Translation Quality Assessment, version 1.0. LT3 Technical Report-LT3 13.02. Retrieved 27 October, 2017 from https://www.lt3.ugent.be/media/uploads/publications/2013/Technical%20 Report%20TQA%20Annotation.pdf.

Frankenberg-Garcia, A. (2015). Training translators to use corpora hands-on: Challenges and reactions by a group of thirteen students at a UK university. *Corpora*, 10(3), 351-380.

Jiménez-Crespo, M. (2009). The Effect of Translation Memory Tools in Translated Web Texts: Evidence from a Comparative Product-Based Study. *Linguistica Antverpiensia*, 8, 213-232.

Kübler, N. (2003). Corpora and LSP translation. In F. Zanettin, S. Bernardini & D. Stewart (eds). *Corpora in Translator Education*. Manchester: St. Jerome Publishing, 25-42.

Kübler, N., Mestivier, A., Pecman, M. & Zimina, M. (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 13èmes Journées internationales d'analyse statistique des données textuelles (JADT 2016)*, Nice, France, 7-10 June 2016, 731-741.

Wilkinson, M. (2010). Quick Corpora Compiling Using Web as Corpus. *Translation Journal*, 14(3). Retrieved January 8, 2016 from http://translationjournal.net/
journal/53corpus.htm.

Zaretskaya, A., Corpas Pastor, A. & Seghiri, M. (2015). Translators' requirements for translation technologies: A user survey. In *Proceedings of the AIET17 Conference New Horizons in Translation and Interpreting Studies*, Malaga, Spain, 29-31 January 2015. Geneva: Tradulex, 247-254.

Text classification for detection of translationese in novice and professional translations

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
e.lapshinova@mx.uni-saarland.de

The aim of the present paper is to compare professional and student translations in terms of register variation, i.e. language variation according to context (Halliday & Hasan 1989; Quirk et al. 1985). We focus on language-dependent register variation in translation, i.e. we prove how well translated texts obey linguistic conventions of both the source and the target languages (English and German in our data) in terms of registers. These conventions are operationalised as particular distributions of lexico-grammatical features according to a given contextual configuration. We focus on translations from English into German and the data at hand includes seven registers of written discourse: political essays, fictional texts, instruction manuals, popular-scientific articles, letters-to-shareholders, prepared political speeches and tourism leaflets. The English originals, professional translations into German and comparable (containing the same registers) German originals were exported from CroCo (Hansen-Schirra et al. 2012), whereas student translations were taken from VARTRA (Lapshinova 2013).

Following the hypotheses of normalisation (translated texts should normalise the linguistic features in order to adapt them to target language conventions) and shining through (linguistic features of the source language may be present in translated texts) – see e.g. Teich (2003), Baker (1996) amongst others – we investigate if register settings of translations correspond to those of the comparable originals (both in the source and the target language). So, linguistic conventions are defined as register profiles on the basis of comparable data in the form of original, non-translated texts in English and German. These register-specific profiles are based on quantitative distributions of features characterising certain registers derived from Systemic Functional Linguistics (SFL, Halliday 2004) and Genre/Register theory (Halliday & Hasan 1989; Biber, 1995). The features represent lexico-grammatical patterns of more abstract concepts, e.g. textual cohesion expressed via pronominal or nominal reference, modality and modal meanings expressed via certain modal verbs. We adopt the feature dataset presented by Lapshinova-Koltunski (2017: 213). These features are content-independent, i.e. they do not contain terminology or keywords.

Our preliminary results demonstrate that both professional and student translations differ from original (source and target) texts in terms of register. Similar observations were made in other studies, such as those by Gellerstam (1986), Baker (1995) and Teich (2003), who show that translations tend to share a set of lexical, syntactic and/or textual features. Pastor et al. (2008) demonstrated this tendency for both professional and student translations. Neumann (2013) analyses an extensive set of linguistic patterns reflecting register variation and demonstrates to what degree translations are adapted to the requirements of different registers, showing how both register and language typology are at work. Several studies, including Ozdowska & Way (2009), Baroni & Bernardini (2006), Kurokawa et al. (2009), Ilisei et al. (2010) and Lembersky et al. (2012), employ computational techniques to investigate these differences quantitatively, mainly applying text classification methods. To our knowledge, none of them use register-related features. The only study that represent linguistic conventions in terms of registers is Lapshinova & Vela (2015). However, the authors compare English-German translation (both human and machine) with non-translated German texts that, as the authors claim, represent target language quality conventions. Their main aim is to show that the usage of translation corpora in machine translation should be treated with caution, as human translations do not necessarily correspond to the quality standards that non-translated texts have.

To analyse register-related differences between English-German human translations and non-translated English and German, we employ text classification with support vector machines (SVM), a technique available in the tool WEKA (Frank et al. 2016). As an input we use matrices containing information on the distribution of the register-related features across various texts in (1) English source texts; (2) German comparable texts; (3) English-to-German translations produced by professionals; (4) English-to-German translations produced by students. We perform several classification tasks using two scenarios: (a) professional translations are used as test data; (b) student translations are used as test data. In both scenarios, we run two classification tasks, where either English originals (EO) or German originals (GO) are used for classifier training, which means that we train register models on original texts and prove if translated texts fit into this model. The performance of classifiers is judged in terms of *f-measure* scores. They are class-specific and indicate the results of automatic assignment of class labels to certain texts. We analyse the results in the following way: if in the scenario where the model was trained on German originals, we observe f-measure over 50% (0.500), translations resemble German comparable texts and, thus, we observe the phenomenon of normalisation. In cases, where f-measure is higher than 50% and English originals were used for training, we observe shining through.

In Table 1, we outline the results sorted according to the registers at hand.

	professional		**student**	
	trained on EO	**trained on GO**	**trained on EO**	**trained on GO**
political essays	0.833	0.439	0.846	0.526
fictional texts	0.343	0.857	0.353	1.000
instructions	0.100	0.480	0.154	0.429
popular-scientific	0.375	0.556	0.571	0.500
letters-to-shareholders	0.000	0.250	0.000	0.333
political speeches	0.258	0.286	0.267	0.364
tourism leaflets	0.625	0.500	0.500	0.500
weighted average	0.403	0.431	0.431	0.477

Table 1. Classification results for both scenarios

Overall, translations do not seem to show clear normalisation or shining through effects, as weighted average is lower than 50% in both cases (0.403 and 0.431 for professional translations and 0.431 and 0.477 for student translations). However, student translations seem to normalise more, as the score for the GO-trained model here is higher than the score for the same model tested on professional translations (0.477 vs. 0.431). At the same time, we see a register-based variation in the results both for professional and student translations. For instance, English seems to 'shine through' in political essays (0.833 and 0.846), whereas fictional texts show obvious normalisation (0.857 and 1.000 respectively). Variation in translations (in terms of shining through and normalisation) is observed for popular-scientific articles and tourism. In the first case, we observe normalisation for professional translations, whereas student translations seem to reveal shining through. In the second case, professional translations seem to be closer to the English sources (show shining through), whereas student translations do not reveal a clear tendency.

In our presentation, we will provide more details on the used set of features, the methods applied and the obtained results including their interpretations.

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: Benjamins, 175-186.

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

Biber, D. (1995). *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Corpas Pastor, G., Mitkov R., Afzal, N. & Garcia Moya, L. (2008). Translation universals: Do they exist? A corpus-based and NLP approach to convergence. In *Proceedings of the LREC 2008 Workshop on Building and Using Comparable Corpora*.

Frank, E., Hall, M. A. & Witten, I. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.).

Halliday, M. (2004). *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M. & Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

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

Ilisei, I., Inkpen, D., Corpas Pastor, G. & Mitkov, R. (2010). Identification of Translationese: A supervised learning approach. In A. Gelbukh (ed.) *Proceedings of CICLing-2010*, LNCS 6008. Heidelberg: Springer, 503-511.

Kurokawa, D., Goutte, C. & Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, Ontario, Canada, 81-88.

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.

Lapshinova-Koltunski, E. (2017). Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. In G. De Sutter, M.-A. Lefer & I. Delaere (eds). *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin: Mouton de Gruyter, 207-234.

Lapshinova-Koltunski, E. & Vela, M. (2015). Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach. In *Proceedings of EMNLP2015 Workshop on Discourse in Machine Translation*, Lisbon, Portugal, 122-131.

Lembersky, G., Ordan, N. & Wintner, S. (2012). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4), 799-825.

Neumann, S. (2013). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: Mouton de Gruyter.

Ozdowska, S. & Way, A. (2009). Optimal bilingual data for French-English PBSMT. In *Proceedings of the EAMT2009 – 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). A comprehensive grammar of the English language. London: Longman.

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

Using corpora for contrastive analysis & translation-oriented ESP teaching

Sara Laviosa
University of Bari Aldo Moro
sara.laviosa@uniba.it

The recent 'multilingual turn' in Applied Linguistics foregrounds 'multilingualism, rather than monolingualism, as the new norm of applied linguistic and sociolinguistic analysis' (May 2014: 1). This paradigm shift entails a re-evaluation of reflexivity in language and intercultural education (cf. Byrd Clark & Dervin 2014). It also entails a rethinking of the transmissionist model of language learning in favour of a model that privileges mutual exchange of knowledge between teacher and students and among students themselves, who are active participants in the development of what is taught, as envisaged by the American educational philosopher John Dewey (1938). The multilingual paradigm has contributed to the re-evaluation of the role played by the L1 in Second Language Acquisition (SLA) (cf. Ellis & Shintani 2014; Ortega 2014), the conceptualization of translanguaging in Bilingual Education (García & Li Wei 2014) and the revival of translation in language teaching (Cook 2010; Laviosa 2014a,b). Engendered by concerns raised in globalization, cosmopolitanism and migration studies (cf. Bielsa 2016; Inghilleri 2017), the multilingual turn is endorsed and promoted by such political forces as the Council of Europe. The pilot extended version of the CEFR illustrative descriptors and the provisional edition of the CEFR companion volume with new descriptors underscore the importance of mediation in honing plurilingual and pluricultural competence. At the core of mediation are different types of translation and translanguaging activities, namely relaying specific information, processing text, explaining data and translating a written text in speech and writing (Council of Europe 2016, 2017). In a similar vein, the new Australian Curriculum for languages adopts an intercultural orientation, and includes translation and interpreting as forms of intercultural mediation involving the analysis and understanding of language and culture as resources for interpreting and shaping meaning in intercultural exchange (ACARA 2014 in Scarino 2016: 480). In unison with the multilingual paradigm are the recommendations made in Ad Hoc Committee Report on Foreign Languages issued by the Modern Language Association of North America (MLA 2007). This programmatic document states that the goal of languages education in the 21st century is to develop translingual and transcultural competence. In contrast to seeking to replicate the competence of an educated native speaker, "[t]he idea of translingual and transcultural competence places value on the ability to operate between languages", and entails the capacity to reflect on the world and on ourselves through the lens of another language and culture (MLA 2007: 3-4). The report also recommends the development of programmes in translation and interpretation because "[t]here is a great unmet demand for translators and interpreters, and translation is an ideal context for developing translingual and transcultural abilities as an organizing principle of the language curriculum" (MLA 2007: 9). Translation is therefore reappraised both as a means of achieving translingual and transcultural competence and a skill in its own right. Against this background, the aim of the present paper is threefold. First, it examines, from a cross-disciplinary perspective, the convergent theoretical underpinnings of the multilingual paradigm. Next, it explores different ways in which, in the specific context of ESP learning in undergraduate and graduate degree programmes, the teacher-scholar and the students-researchers engage collaboratively with task-based and translation-oriented activities involving cross-lingual and cross-cultural analyses of multilingual corpora, drawing on the methodology and empirical insights of corpus-based contrastive and translation studies. Third, it reports on the rationale, methods and findings of an exemplary classroom-based observation study of a pedagogy that combines translanguaging and translation, and draws on Tim Johns' Data-Driven Learning approach (Johns 1991). The study was carried out by the author as a participant observer of her own class during a PhD seminar on "Translanguaging and Translator Training" given on 27.03.2015 at IUSLIT, University of Trieste. The learning objectives of the pedagogic unit were: a) to introduce bilingual reference corpora as computer-aided translation tools; b) to engage in translanguaging activities for translation purposes in order to raise cross-lingual and cross-cultural awareness, in line with the principles and methods of corpus-based contrastive studies. The resources and materials included two comparable reference corpora, i.e. the *British National Corpus* (BNC) and *Corpus di Italiano Scritto* (CORIS), and an article on the Greek crisis published in *The Economist* on 25.02.2015. The activities consisted of an introduction to corpora and translanguaging, followed by corpus-based translanguaging aimed at producing an accurate, fluent and pragmatically unmarked Italian translation of the English term *austerity* in the subject-specific discourse of political economics, particularly in the text genre of press articles. It is fair to affirm that, though small-scale and merely observational, this study shows that, in an ESP learning environment where ELF is used in a monolingual-endolingual mode, translating through corpus-based translanguaging has the distinctive potential to achieve convergent goals that are congruous with the tenets underpinning the multilingual paradigm. It valorizes the bilingual repertoire of learners and teachers, it benefits the linguistic processing of subject content, it enhances multilingual reasoning, and contributes to the mediation and construction of knowledge in both the L1 and the L2. These achievements are fundamental for developing pluriliteracy abilities, and enabling students to become members of the international scientific and academic discourse community. Looking to the future, the corpus-based pedagogy adopted in this study would benefit enormously from the development of corpus-informed tools and resources for ESP learning through translation (e.g. the MUST corpus at the Centre for English Corpus Linguistics of the University of Louvain) as well as further ethnographic investigations carried out in a variety of educational settings and in a wide array of languages.

References

Bielsa, E. (2016). *Cosmopolitanism and Translation: Investigations into the Experience—of the Foreign*. London & New York: Routledge.

Byrd Clark, J. S. & Dervin, F. (eds). (2014). *Reflexivity in Language and Intercultural Education: Rethinking Multilingualism and Interculturality.* London & New York: Routledge.

Cook, G. (2010). *Translation in Language Teaching. An Argument for Reassessment*. Cambridge: Cambridge University Press.

Council of Europe (2016). *CEFR Illustrative Descriptors – Extended Version 2016*. Pilot version for consultation. Strasbourg: Language Policy Unit [online]. Available at https://mycloud.coe.int/index.php/s/VLAnKuMxDDsHK03 [Accessed 21 December 2017].

Council of Europe (2017). *CEFR Companion Volume with New Descriptors*. Provisional edition [online]. Available at https://www.coe.int/en/web/common-european-framework-reference-languages [Accessed 21 December 2017].

Dewey, J. (1938). *Experience and Education*. New York: Touchstone.

Ellis, R. & Shintani, N. (2014). *Exploring Language Pedagogy through Second Language Acquisition Research*. London & New York: Routledge.

García, O. & Li Wei (2104). *Translanguaging. Language, Bilingualism and Education*. London: Palgrave Macmillan.

Inghilleri, M. (2017). *Translation and Migration*. London & New York: Routledge.

Johns, T. (1991). Should You Be Persuaded: Two Samples of Data-Driven Learning Materials. *Classroom Concordancing. ELR Journal* 4, 1-16.

Laviosa, S. (2014a). *Translation and Language Education: Pedagogic Approaches Explored*. London & New York: Routledge.

Laviosa, S. (ed.) (2014b). Translation in the Language Classroom: Theory, research and practice. Special Issue of *The Interpreter and Translator Trainer* 8(1).

May, S. (2014). Introducing the 'Multilingual Turn'. In S. May (ed.) *The Multilingual Turn: Implications for SLA, TESOL and Bilingual Education*. London & New York: Routledge, 1-6.

MLA Ad Hoc Committee on Foreign Languages (2007). Foreign Languages and Higher Education: New Structures for a Changed World. 1-12. Available at www.mla.org/flreport [Accessed 21 December 2017].

Ortega, L. (2014). Ways Forward for a Bi/Multilingual Turn in SLA. In J. Conteh & G. Meier (eds). *The Multilingual Turn in Languages Education: Opportunities and Challenges*. Bristol: Multilingual Matters, 32-53.

Scarino, A. (2016). Reconceptualising Translation as Intercultural Mediation: A renewed place in language learning. *Perspectives: Studies in Translatology*. *Translation as Intercultural Mediation* 24(3), 470-485.

How have interpreting norms changed at premier press conferences in China? A corpus-based study

Nannan Liu
The University of Hong Kong
nnl93@hku.hk

This research intends to uncover the diachronic change of interpreting norms at premier press conferences in China. Corpus linguistics represents a powerful methodology for interpreting researchers to investigate a huge body of texts and unravel patterns of language use in interpreting scenarios. Current researches, however, have been limited to interrogation of words, grammar and interpreting strategies (see for example Sergio & Falbo 2012; Hu & Tao 2013; Wang 2012). Little attention has been directed to discourse and historical linguistics (Barðdal 2002). While researchers capitalize on corpus technology, regularities of interpreted language are, more often than not, attributed to contrast of source and target languages and nature of conference interpreting, rather to the context and norms of the mediated events (Stewart 2000; Xiao 2012; Ramon 2015). A gap remains to locate the norms and their possible changes with recourse to corpus linguistics.

To this end, this research built a bilingual Corpus of Interpreted Premier Press Conferences (SCIPPC). It collected a repository of texts from 6 annual press conferences at China's parliament, and was complemented with a database of video footages of the events. The data ranges from 2003 to 2005, and 2013 to 2015 with input by 5 different interpreters. Official "transcripts", which are reviewed by government censors, of both source and target texts were gathered, revised, and annotated to reproduce the linguistic, paralinguistic, and extra-linguistic reality of the events. SCIPPC was annotated with XML and processed with programming language Python. Facilitated by Natural Language Toolkit, this research was greatly expedited with regard to data cleaning, mark-up and queries, which are often considered most onerous in corpus linguistics (Zhang 2012).

This research draws its descriptive paradigm from Toury's notion of norms (Toury 2012; Pym et al. 2008). Toury argues that translators and interpreters always make a choice between two extreme orientations of leaning on the original norms and target norms (2012: 79). Any interpreting utterance would be located on the continuum of adherence to source norms (interference) and target norms (standardization) (Toury 2012: 303-315). The research question is thus dissected to 1) how much the interpreted English in SCIPPC is interfered by the original, and 2) how much it is standardized according to target norms.

SCIPPC was part-of-speech tagged and parsed according to turn-taking in the authentic settings. Time lag between source and target utterances was recorded. Three types of parameters are marked and investigated in the SCIPPC: linguistic, paralinguistic and extralinguistic. Linguistic patterns include: the use of pronouns ("we", "us", "our" in particular), keywords in Chinese cultural-political universe, typical syntactical structure (topic plus comment), discourse anaphors, shifts between the source and target, metaphors and government censored parts. Regular para- and extra-linguistic phenomena *in situ*, such as turn-taking, interaction, interruption, eye contact and event profiles were also documented. The interpreted language was treated on its own right, and the relationship of the source vis-à-vis target texts was examined through correlation tests of lexical, syntactical and semantic shifts.

Preliminary investigation of SCIPPC and the video database suggests that the interpreted language displays less interference by the source and growing standardization. Interpreting norms are shifting from heavy dependence on the source (adequacy) to more emphasis on acceptability. In the interpreted language, this is seen from 1) decreasing carryover of collective "we", "our", and "us", but more target-oriented use of "China"; 2) increasing use of simple words and shorter sentences; 3) fewer repetitions. In examining the relationship between source and target, interference is found to be less marked considering there are 1) more intelligible translation of keywords and involving figures of speech; 2) less carryover of typical Chinese syntactic structures; 3) less frequent use of addition and revision, but much more omission.

SCIPCC also showcases more interaction, reference among the participants, less applause and better coordination between speakers and interpreters. The case of censorship, however, tells a different story. Censorship not only increases over time in SCIPCC, but also becomes more sophisticated, which might point to changing political norms in which interpreting operates.

Social-cultural variables such as more interaction between China and the rest of the world, prevalence of media, institutional building of the Chinese government, and popularity of English as a foreign language and Chinese as a foreign language, contribute to the changing norms of interpreting at these press conferences.

This research bridges the gap by using authentic data to analyse how interpreting norms change. It represents an empirical attempt to test Toury's notion of norms and laws of interference and growing standardization. It employs Python, a powerful but rare find tool in today's corpus-based interpreting studies, to significantly assist processing of language data. It also offers new perspectives to China studies as it describes how the monumental changes in contemporary China are shaping its language use and interpreting practice. It holds implications for corpus-based interpreting studies and natural language processing.

References

Baker, M. (1993). 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: John Benjamins, 233-250.

Baker, M. (1996). Corpora in Translation Studies: The Challenges that Lie ahead. In Harold Somers (eds). *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 175-186.

Barðdal, J. (2002). "A Crash Course in Corpus Linguistics". Retrieved from http://www.cas.unt.edu/~jbarddal/corpus.html. Accessed on 20-09-2017.

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

Biber, D. & Gray, B. Being Specific about Historical Change: the Influence of Sub-Register. *Journal of English Linguistics* 41(2), 104-134.

Chen, L. (2007). *Cultural Context and Political Discourse: A Discourse Analysis of Government Press Conference*. Beijing: China Radio Film and TV Press.

Hu, K. & Qing, T. (2013). The Chinese-English Conference Interpreting Corpus: Uses and Limitations. *Meta: Translators' Journal* 58(3), 626-642.

Kennedy, G. (2000). *An Introduction to Corpus Linguistics*. Beijing: Foreign Language Teaching and Research Press.

Liang, M., Li, W. & Xu, J. ((2010). *Using Corpora: A Practical Coursebook*. Beijing: Foreign Language Teaching and Research Press.

Oaks, M. P. & Ji, M. (eds). (2012). *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research*. Philadelphia: John Benjamins Publishing Company.

Pym, A., Shlesinger, M. & Simeoni, D. (2008). *Beyond Descriptive Translation Studies*. Philadelphia: John Benjamins Publishing Company.

Ramon, N. (2015). Comparing Original and Translated Spanish: A Corpus-based Analysis of Adjective Position. *Babel* 61(4), 527-551.

Sergio, F. S. & Falbo, C. (eds). (2012). *Breaking Ground in Corpus-based Interpreting Studies*. Bern: Peter Lang.

Shlesinger, M. (1989a). Extending the Theory of Translation to Interpretation: Norm as a Case in Point. *Target* 1(1), 111-115.

Shlesinger, Miriam (1989b). *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum*. MA thesis, Tel Aviv University.

Stewart, D. (2000). Conventionality, creativity and translated text: The implications of electronic corpora in translation. In M. Olohan (ed.) *Intercultural Faultlines: Research Models in Translation Studies*. Manchester: St Jerome Publishing, 79-91.

Toury, G. (2012). *Descriptive Translation Studies – and Beyond*. Philadelphia: John Benjamins Publishing Company.

Xiao, R. (2010). How Different is Translated Chinese from Native Chinese? A Corpus-based Study of Translation Universals. *International Journal of Corpus Linguistics* 15(1), 5-35.

Wang, B. (2012). Corpus-based Interpreting Studies – A Breakthrough in the Research of Interpreting Products. *Foreign Languages in China* 9(3), 94-100.

Wynne, Martin (ed.) (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. Retrieved from http://ota.ox.ac.uk/documents/
creating/dlc/ (Accessed 22-10-2017).

Zhang, W. (2012). Interpreting Corpus and Relevant Researches in the Last Decade: Present Conditions and Oncoming Trends. *Journal of Zhejiang University (Humanities and Social Sciences)* 42(2), 193-205.

Using non-standard word order you should! A corpus-based approach to avoiding standardized word order in translated French

Rudy Loock
Université de Lille, UMR « Savoirs, Textes, Langage » du CNRS
rudy.loock@univ-lille.fr

The aim of our presentation is to show how the use of different electronic corpora in translation training (comparable and parallel, containing original and/or translated texts, learner and professional) can help sensitize students to the importance of word order for the idiomaticity/naturalness of their translations.

Our presentation will focus on non-standard (or marked) word order in English and French: clefting (1), pseudo-clefting (2), left/right dislocation (3), topicalization (4), as opposed to the canonical, standard SVO word order – subject-verb-object (5):

(1) It is a new computer that the president has bought. / *C'est un nouvel ordinateur que la présidente a acheté*.
(2) What the president has bought is a new computer. / *Ce que la présidente a acheté, c'est un nouvel ordinateur*.
(3) a. The president, she has bought a new computer. / *La présidente, elle a acheté un nouvel ordinateur.*
b. She has bought a new computer, the president. / *Elle a acheté un nouvel ordinateur, la présidente.*
(4) A new computer the president has bought. / *Un nouvel ordinateur la présidente a acheté.*
(5) The president has bought a new computer. / *La présidente a acheté un nouvel ordinateur.*

Note that other syntactic constructions like the use of the passive voice, existential and presentational constructions, locative inversion, or extraposition are sometimes listed as non-standard constructions, but we are focusing here on the constructions listed in (1)-(4).

While the non-standard syntactic organizations in (1)-(4) exist both in English and in French, the discourse constraints for them to appear felicitously and their frequencies are not similar (e.g. Birner & Ward 1998; Lambrecht 1994; Carter Thomas 2002). For instance, the French language shows much more use of non-standard word orders, as opposed to the English language, where the SVO word order is very frequent. For both languages, the choice between the different "allostructures" (1)-(5) is never random.

The starting point of this presentation is the observation that students, even advanced students, translating from English into French have a tendency to overuse the SVO, standard word order, which can be interpreted as being due to both interference from the source language (English) and also perhaps the tendency to "normalize" the target language, as put forward by Baker (1996: 184): "*[m]arked, rather than ungrammatical structures, are also often "normalised" in translation*"). Our starting point will be illustrated thanks to the analysis of a learner corpus of advanced students' English>French translation tasks, all students being native speakers of French and the type of texts in the corpus being press texts.

Thanks to the use of a comparable corpus of English-French original press texts, we will show that (i) English and French do not use the types of non-standard, marked syntactic organizations in (1)-(4) with the same frequencies, and (ii) the students' translations thus show a difference with original French, possibly leading to a lack of idiomaticity.

Our discussion will turn to the consequences of such results for the teaching of translation, and will be illustrated by the methodology used in a comparative grammar class within a master's program in our university, where the use of both comparable and parallel corpora, the latter containing translations by professional translators, aims to improve the naturalness of translated texts by taking account language use, beyond grammatical correctness (Loock 2016, 2017).

Finally, our presentation will analyze a corpus of French texts translated thanks to (neural) machine translation, the English original texts being extracted from our comparable corpus, in order to show that the ability of dealing with word order phenomena is (still) the prerogative of human translators. This type of corpus is analyzed in class with students in order to show them the extent to which machine translation systems can take into account language use, the final aim being to make students aware of human translators' added value over machine translation, a real challenge for translators-to-be.

Our presentation is therefore based on the analysis of four different 'do-it-yourself' (DIY) corpora, for both inter-language/cross-linguistic and intra-language comparisons:

(i) a learner corpus of French texts translated from English by advanced translation students;
(ii) a French-English comparable corpus of original texts;
(iii) a corpus of French machine translated texts from English;
(iv) a parallel corpus of French texts translated from English by professional translators.

In addition to providing results, our presentation will discuss methodological issues for corpus studies dealing with such linguistic phenomena (corpus, size, representativeness, manual vs. automatic search).

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 Publishing Company, 175-186.

Birner, B. J. & Ward, G. (1998). *Information Status and Noncanonical Word Order in English*. Amsterdam: John Benjamins Publishing Company.

Carter-Thomas, S. (2002). Theme and Information structure in French and English: A contrastive study of journalistic clefts. *14th Euro-International Functional Systemic Workshop*, Lisbon University, Lisbon, Portugal.

Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents*. Cambridge: Cambridge University Press.

Loock, R. (2016). *La Traductologie de corpus*. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Loock, R. (2017). Because Grammatically Correct is not Enough: Grammatical Naturalness in the Target Language as the Icing on the Cake for Future Translators. International conference *Quelle formation grammaticale pour le futur traducteur?*, 8-9 March 2017, Mons, Belgium.

A corpus-based study on lexical simplification across interpreting types

Qianxi Lv, Junying Liang
Zhejiang University
vera_lv52e@126.com, jyleung@zju.edu.cn

Among the various modes of interpreting, simultaneous interpreting (SI) has been addressed by different authors as a ‘complex’ (De Groot 2000) , ‘extreme condition’ (Meuleman & Van Besien 2009; Obler 2012) of cognitive tasks. Consecutive interpreters, on the other hand, do not have to share processing capacity between tasks under a high cognitive load; nor are there problems arising from ‘an accumulation of tasks under the pressure of time resulting in capacity requirement peaks’ (Gile 2009). Furthermore, in consecutive interpreting (CI), the presumably higher cost of speech production in the B language could be accommodated in the self-paced reformulation stage. Given that SI exerts great cognitive demand, it makes sense to posit that the output of SI may be more lexically compromised than that of CI.

The bulk of the research addressing the issue of interpreting has stressed the varying cognitive demand and processes involved in different modes of interpreting. The inconsistency of corpus-driven interpreting studies also reveals the possible existence of a ‘modality effect’ (the differences of interpreting types) in the lexical context. Thus, a quantitative investigation into different types of interpreting is needed to discern the specific features that could set each of them apart.

The present study serves as a product-oriented study comparing the cognitive demand of SI versus CI, with a self-made inter-model corpus of transcribed interpretation, translated speech and original speech texts. We performed the calculations using uniform methods across modes of interpreting, and controlled for possible confounding textual factors.

The corpus was constructed with transcribed real-world materials for four sub-corpora of similar size and a total running words of 28,1960, namely, 1) a CI corpus consisting of the interpretation of press conferences of the National People’s Congress from 2009 to 2016 in China; 2) a corpus of SI texts made up of 21 interpretations of keynote speeches recorded at the Boao Forum of Asia, Davos Forum from 2009 to 2016, as well as BRICs summits, sessions of the U.N. General Assembly, and China-ASEAN conferences; 3) a read-out translated speech corpus (Tr-sp) of recorded government work reports from 2009 to 2016; and 4) an non-interpreted, original English speech corpus (Or-sp) of State of the Union Messages from 2009 to 2016.

The results of the present study are the very first observations of their kind suggesting that different modes of interpreting affect lexical traits of the output in a quantitative context. While previous studies on interpreting studies generally regard SI as an extreme situation of multitasking with the highest cognitive load, our findings evidently show that CI imposes heavier cognitive demands and hence yields more lexically simplified output. CI is more repetitive, less informative, and less sophisticated than the original English speeches and read-out translations. Conversely, in SI, although it is more lexically repetitive than non-interpreted discourse, the level of informativeness and sophistication is even higher than the original speeches, as it is diluted with higher Lexical Density and lower Core Vocabulary Coverage. This result of SI corroborates with previous observation and research (Chachibaia & Colenso 1998).

This pattern of results implies that the cognitive load of consecutive interpreting, if not higher, may be as high as that of simultaneous interpreting. The load is exerted both on the maintenance component and on the executive control mechanism of working memory, as established in Baddeley’s working memory model (Baddeley & Hitch 1974). On the one hand, the short-term memory effort corresponds to the storage component of WM, with an emphasis on its transient nature. On the other hand, the executive function consists of the Coordination of all other processing efforts in these models. According to Gile (2009), all types of effort are competing for limited processing resources, and one type of effort may suffer due to the rising load on another.

The information to be maintained in CI is inherently larger in volume compared to SI. It is apparent that the storage component taxes CI interpreters much more than SI interpreters and thus endangers interpreters' ability to keep their attention oriented towards encoding information for the output.

In terms of the executive functions, we argue, despite the fact that the CIs are free to allocate the processing capacity at their own pace in Phase Two (Gile 2009), pressure on the Coordination component remain. The demands of capacity on CI are diversified – constantly switching between two languages and meanwhile inhibiting one while rendering another (simultaneously reading notes which are at least partially in the source language and producing in the target language). The increment in these efforts is also possibly due to anticipation during CI to ensure fluent, appropriate interpreting. When strongly constrained by time in such situations as press conferences, a division of attention between delivering some part of the speech and reading what comes next in the notes seems to be intrinsic (Gile 2015).

The underlying reasons could be the different cognitive demand between SI and CI, and here we propose a revised effort model adapted from Gile (2017). The modification has been made to address both the storage and executive function components of the WM model (Baddeley & Hitch 1974) based on the assumption that both task components deplete the same reservoir of attentional resources (Christoffels & Groot 2004). We propose to specify the distinctive executive control mechanisms for SI and the reformulation phase of CI respectively according to the evidence from behavioral experiments.

References

Baddeley, A. D. & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation - Advances in Research and Theory* 8(C), 47-89.

Chachibaia, N. G. & Colenso, M. R. (1998). Simultaneous interpreting of a scientific discussion. *Perspectives* 6(2), 217-224.

Christoffels, I. K. & de Groot, A. M. B. (2004). Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition* 7(3), 227-240.

De Groot, A. M. B. (2000). A Complex-skill Approach to Translation and Interpreting. In S. Tirkkonen-Condit & R. Jääskeläinen (eds). *Tapping and Mapping the Processes of Translation and Interpreting Outlooks on empirical research*. Amsterdam & Philadelphia: John Benjamins, 52-70.

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

Gile, D. (2015). Testing the effort models' tightrope hypothesis in simultaneous interpreting – a contribution. *Journal of Linguistics* 35(2), 590-647.

Gile, D. (2017). The effort models and gravitational model- clarifications and update [PPT]. Retrieved from http://www.cirinandgile.com/powerpoint/The-Effort-Models -and-Gravitational-Model-Clarifications-and-update.pdf.

Meuleman, C. & Van Besien, F. (2009). Coping with extreme speech conditions in simultaneous interpreting. *Interpreting* 11(1), 20-34.

Obler, L. K. (2012). Conference interpreting as extreme language use. *International Journal of Bilingualism* 16(2), 177-182.

Norms and gender in simultaneous interpreting: A study of self-repairs

Cédric Magnifico, Bart Defrancq
Ghent University
cedric.magnifico@ugent.be, bart.defrancq@ugent.be

Introduction

This paper focuses on a possible gendered approach of norms in simultaneous interpreting. Interpreting is subject to various types of norms: translational norms (Schlesinger 1989; Harris 1990; Toury 1995), norms acquired through training and professional experience (Schjoldager 1995; Duflou 2016), and norms which derive from the expectations of the different actors taking part in the interpreting process (Bülher 1986; Kopczynski 1994; Kurz 2000; Garzone 2002). However, interpreters appear to breach some of these norms in specific settings to accommodate other needs (Jansen 1992; Wadensjö 1998; Monacelli 2009).

From a gender perspective, research conducted in the field of linguistics shows that men and women do not value the same norms (Labov 1966, 1990; Trudgill 1972). As interpreting is a specific kind of language (Gile 1995), we can expect gender differences in the approach towards norms. Recent studies have revealed that gender influences interpreters' expectations (Pöchhacker & Zwischenberger 2010) and their interpreting strategies (Cecot 2001; Magnifico & Defrancq 2016, 2017). Magnifico & Defrancq (2017) observe that female interpreters tend to translate face-threatening acts more straightforwardly than their male colleagues and suggest that it could be due to a gender-driven choice of the followed norms.

Although research on norms has gained in interest over the last decades, this field is not well researched (Gile 1998) and lacks empirical studies focusing on the norms applied by interpreters in authentic settings. In this respect, a corpus-based study can identify recurring patterns in the use of norms by interpreters themselves. But can norms be operationalized to suit a corpus-based analysis? We will argue that self-repairs, which are researchable textual elements, manifest the norms expressed by interpreters. Drawing on Levelt's (1983) framework, we define self-repairs as corrections made by the interpreter without any external stimulus occurring in three stages: (a) the utterance (reparandum), (b) the interruption of the flow of speech, with or without an editing term, and (c) the repair proper, i.e. the new utterance.

Considering the aforementioned literature, we will examine the following research questions:

- Is there a gender difference in the number of self-repairs produced by male and female interpreters?
- Does gender influence the way male and female interpreters self-repair their output?

Data

The EPICG (European Parliament Interpreting Corpus Ghent) corpus is being compiled at Ghent University following the Valibel norms (Bachy et al. 2007) and using the method described in Bernardini et al. (2018). Speeches held at the European Parliament during plenary sessions and their interpretations in different languages are transcribed from video footages available on the website of the European Parliament. The transcriptions reflect a number oral features, such as false starts, repetitions, etc. The current corpus comprises about 220,000 tokens in 9 language combinations: French / Dutch / English / German. For the present study, we used the 2008 sub-corpus which contains 193,000 words and 39 speeches in French and 39 interpretations, both in English and in Dutch.

Methodology

We manually examined the English and Dutch texts to identify all the occurrences where the interpreter breaks the flow of speech. We then selected all the self-repairs meeting the definition outlined above. When it appeared that the interpreter made a false start, i.e. repeated the same syllable or word, we discarded the occurrence. We further classified the self-repairs according to the strategy used by the interpreter: self-repairs without and self-repairs with editing terms. We finally conducted a chi-square test to see whether the gender differences observed in the number of self-repairs and in the self-repair strategy were significant.

Results

This empirical corpus-based study reached interesting results. The first research question is confirmed: it appears that female interpreters significantly self-repair more utterances than male interpreters. The second research question, however, yielded surprising results as the pattern differs according to the language pair. In the French-English language combination, female interpreters use significantly more editing terms than their male colleagues. In the French-Dutch language pair, no difference is observed in the use of editing terms. However, we noticed that Dutch female interpreters use apologies as editing terms, three times more than their male colleagues. Nevertheless, the number of occurrences is far too limited to draw conclusions in this respect. Incidentally, we also observed that male and female interpreters not only self-repair erroneous, but correct utterances as well.

As self-repairs are an operationalization of the norm expressed by interpreters, we can conclude that female interpreters take a more normative approach in their interpretations than male interpreters. Further research with a larger corpus and more language pairs is needed to confirm these trends and gain a better insight in the field of interpreting norms.

References

Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.-C. & Wilmet, R. (2007). *Conventions de transcription régissant les corpus de la banque de données VALIBEL*. https://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.PDF (accessed 15 October 2013)

Bernardini, S., Ferraresi, A., Russo, M., Collard, C. & Defrancq, B. (2018). Building Interpreting and Intermodal Corpora: A *How-to* for a Formidable Task. In C. Bendazzoli, M. Russo & B. Defrancq (eds). *Making way in corpus-based interpreting studies* (Vol. 1). Singapore: Springer, 21-42.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua* 5(4), 231-235.

Duflou, V. (2016). *Be(com)ing a conference interpreter: an ethnography of EU interpreters as a professional community*. Amsterdam: John Benjamins.

Gile, D. (1995). *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam: Benjamins.

Gile, D. (1998). Norms in Research on Conference Interpreting: A Response to Theo Hermans and Gideon Toury. *Current Issues In Language and Society 5*, 1(2), 99-106.

Harris, B. (1990). Norms in Interpretation. *Target*, 2(1), 115-119.

Jansen, P. (1992). The Role of the Interpreter in Dutch Courtroom Interaction: the Impact of the Situation on Translational Norms. In P. Jansen (ed.) *Selected Papers of the CERA Research Seminars in Translation Studies 1992-1993*. Katholieke Universiteit Leuven, 133-155.

Kopczynski, A. (1994). Quality in conference interpreting: Some pragmatic problems. In S. Lambert & B. Moser-Mercer (eds). *Bridging the Gap: Empirical research in simultaneous interpretation*. Amsterdam: John Benjamins, 87-99.

Kurz, I. (2000). Conference Interpreting: Quality in the Ears of the User. *Meta* 46(2), 394-409.

Labov, W. (1966). *The Social Stratification of English in New York City*. Exi: Center for Applied Linguistics.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2(2), 205-254.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition 14*, 41-104.

Magnifico, C. & Defrancq, B. (2016). Impoliteness in interpreting: a question of gender? *Translation And Interpreting* 8(2), 26-45.

Magnifico, C. & Defrancq, B. (2017). Hedges in conference interpreting: The role of gender. *Interpreting* 19(1), 21-46.

Schjoldager, A. (1995). An Exploratory Study of Translational Norms in Simultaneous Interpreting: Methodological Reflections. *Hermes, Journal of Linguistics* 8(14), 65-87.

Shlesinger, M. (1989). Extending the Theory of Translation to Interpretation: Norms as a case in point. *Target* 1(1), 111-115.

Toury, G. (1995). *Descriptive Translation Studies and beyond.* Amsterdam: John Benjamins.

Trudgill, P. (1972). Seks, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1(2), 179-195.

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

The gravitational pull of diminutives in Catalan translated and non-translated texts

Josep Marco, Ulrike Oster
Universitat Jaume I
jmarco@uji.es, oster@uji.es

In the present paper, Halverson's gravitational pull hypothesis (GPH) (2003: 197-198) will be tested in connection with diminutive suffixes, which may be said to be fairly typical of Catalan (and other Romance languages) but are not so productive in Germanic languages, although there may be significant differences among them.

Halverson's GPH is an attempt to show that "the various lexical/semantic patterns that have been subsumed under the heading 'translation universals' may be explained with reference to general characteristics of human cognition". It both pertains to a higher level of generalisation than the various hypotheses subsumed under the heading of translation universals and purports to provide a cognitive explanation for them.

Along the lines laid by Halverson (2009: 93-102), Hareide (2013) put the GPH to the test. She used two parallel corpora (English-Spanish and Norwegian-Spanish) which were assumed to be comparable in all relevant respects. The reason for having two parallel corpora with Spanish as target language was that the whole study pivoted around two grammatical structures which may be said to be unique items for Norwegian-Spanish but not for English-Spanish. Hareide (2013: 36 and ff.) draws on Halverson (2010) to formalise the GPH as follows. There are three potential causes of translational effects: patterns of 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 present study – as in Hareide's – two parallel corpora (English-Catalan, German-Catalan) will be used which can be regarded as comparable in all relevant respects. These two parallel corpora belong to the more general COVALT (Valencian Corpus of Translated Literature) corpus, which also includes Catalan translations from French, and Spanish translations from the three source languages mentioned, as well as two non-translated components (in Catalan and Spanish). We will draw on Catalan non-translations for comparison with translations from English and German.

The reason underlying the choice of diminutives as an indicator for testing the GPH is that diminutives may be regarded as unique items for the English-Catalan language pair but not for the German-Catalan pair, as there are at least two diminutive suffixes in German, *-chen* and *-lein*, which are very frequent and productive. If the notion of uniqueness in this respect appears problematic (as it may well do, since English *does* possess diminutive suffixes, such as *-let* in *piglet*, although they are not very productive), the alternative, more cline-like notion of *degree of overlap* may be posited. Thus, German and Catalan would show a higher degree of overlap than English and Catalan. On this basis, three hypotheses are formulated:

1. Diminutive suffixes will be over-represented (factor 1) or under-represented (factor 3) in translations from English into Catalan, as compared to Catalan non-translations;
2. Diminutive suffixes will be over-represented (factors 1, 2 and 3) in translations from German into Catalan, as compared to Catalan non-translations;
3. Diminutive suffixes will be significantly more frequent in translations from German into Catalan, where overlapping structures exist, than in translations from English into Catalan, where this is much less the case (factors 2 and 3).

Drawing on two Catalan grammars (Cabré 2002 and AVL 2006), three diminutive suffixes were singled out for analysis: *-et*, *-ó* and *-iu* – including their gender and number inflections. These suffixes may be applied to nouns and adjectives. In what follows some preliminary results will be presented as regards suffix *-et* applied to nouns. Data was retrieved from the corpora with the Corpus Query Processor (CQP), and query matches had to be manually sifted, as there are lexical items ending in *-et* which are not diminutives at all or do not qualify as proper diminutives because they have been lexicalised and have a separate dictionary entry. Since the number of query matches was too high for manual analysis (e.g. over 7,000 cases in non-translations), results were randomly thinned for the three components. Thus, the actual number of concordance lines analysed was 500 for translations from English, 500 for translations from German and 700 for non-translations. The three components under comparison are different in size, so frequencies of occurrence were normalised to matches per 1,000 words. Results are shown in Table 1. Finally, log-likelihood was applied to test for statistical significance. Differences between translations from English and Catalan non-translations are statistically highly significant (LL 146.91, p<0.0001); differences between translations from German and non-translations are also significant (LL 10.46, p<0.01), although less so than in translations from English; and differences between translations from German and translations from English are significant too (LL 37.58, p<0.0001), their significance coming somewhere in-between the two significance values just mentioned. The diminutive suffix *–et*, then, is under-represented in translations, whether from English or German, and it is under-represented in translations from English when compared to translations from German. This means that, as regards hypothesis 1, factor 3 (akin to the Unique Items Hypothesis) overrules factor 1. Hypothesis 2 is disproved, as all three factors are predicted to pull in the direction of over-representation but, even so, under-representation occurs. Finally, hypothesis 3 is confirmed.

	Total number of words	Query matches	Proper diminutives (out of 500/700)	Estimated frequency of diminutives in the whole corpus	Normalised frequency per 1,000 words
Translations from German	604,966	3,339	121	808.04	1.34
Translations from English	1,343,631	5,405	126	1362.06	1.01
Non-translations	1,551,521	7,445	222	2361.13	1.52

Table 1. Frequency of occurrence of the diminutive suffix *-et* in Catalan translations from English and German and in Catalan non-translations (COVALT corpus)

The full study will include a quantitative analysis, similar to the one just presented, of the other two suffixes mentioned (*-ó* and *-iu*) as well as a qualitative analysis of the bilingual concordances in order to determine to what extent the differences observed between translations from German and translations from English are due to the occurrence of diminutive suffixes in German texts.

References

Acadèmia Valenciana de la Llengua. (2006). *Gramàtica normativa valenciana*. Valencia: Publicacions de l'Acadèmia Valenciana de la Llengua.

Cabré, M.T. (2002). La derivació. In J. Solà, M. R. Lloret, J. Mascaró & M. Pérez Saldanya (eds). *Gramàtica del català contemporani* [Grammar of contemporary Catalan], *Introducció. Fonètica i fonologia. Morfologia* [Introduction. Phonetics and Phonology. Morphology]. Barcelona: Empúries, 731-775.

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

Halverson, S. (2009). Elements of doctoral training: The logic of the research process, research design, and the evaluation of research quality. *The Interpreter and Translator Trainer* 3(1), 79-106.

Halverson, S. (2010). Cognitive translation studies: developments in theory and method. In G. M. Shreve & E. Angelone (eds). *Translation and Cognition*. Amsterdam & Philadelphia: John Benjamins, 349-369.

Hareide, L. (2013). *Testing the Gravitational Pull Hypothesis in translation. A corpus-based study of the gerund in translated Spanish*. Bergen: University of Bergen (PhD dissertation presented in May 2014).

Multifunctionality of evidential expressions in discourse: A contrastive study of evidential values and variation in discourse domains and construction types

Juana I. Marín-Arrese
Universidad Complutense de Madrid
juana@filol.ucm.es

This paper examines the phenomenon of multifunctionality of evidential expressions in unscripted conversation and in journalistic discourse in English and Spanish. Evidentials have been characterized as primarily indicating the source of information and the evidence on the basis of which the speaker feels entitled to make a claim (Anderson 1986; Aikhenvald 2004). A broader conception of evidentiality includes both the source of information and an estimation of its reliability, as Chafe (1986) posited in his seminal publication (cf. Chafe & Nickolls 1986). It has been argued that the different values of evidentiality are typically associated with different degrees of reliability of the source and mode of access to the evidence, and thus also of hearers' perception of degrees of speaker commitment to the proposition, and hearers' potential acceptance of the validity of the communicated information (Fitneva 2001; Papafragou et al. 2007; Marín-Arrese 2011). From a cognitive-functional perspective, evidentiality is conceived as a subdomain of the conceptual domain of epistemicity, in providing 'epistemic justification' for a proposition (Boye 2012).

Within evidential systems there is a basic distinction between direct (sensory) or indirect (inferential, reportative) modes of access to knowledge, the latter involving mediation by higher-level cognition in the case of inference or mediation through other individuals in the case of report (Langacker 2017). The values 'direct', 'indirect inferential', and 'indirect reportative' evidentiality are typically expressed by particular evidential markers (Diewald & Smirnova 2010). There are, however, a number of evidential expressions in English and Spanish, as in other languages, that are synchronically polyfunctional, which would seem to point to certain bidirectional connecting links between specific subspaces or notional regions within the semantic map of epistemicity (cf. Boye 2012). This paper focuses on the link between direct and indirect justification for a proposition within the domain of evidentiality, as well as on inter-subspace extensions, that is, on multifunctionality involving 'indirect inferential' and 'indirect reportative' values within the subspace of indirect evidentiality.

The paper addresses the following research questions: (i) whether there is a pattern of preferences for the various expressions of evidentiality and their evidential values correlating with discourse-domains and genres, and across languages; (ii) whether the occurrence of particular values of multifunctional evidential expressions may be associated with variation in discourse domains and genres; and (iii) the degree to which certain evidential values (inferential vs. reportative) are characteristic of particular evidential constructions in the two languages.

The paper presents results of a contrastive corpus-based study of the expression of evidentiality and multifunctionality with core evidential expressions (verbs and sentence adverbs). The data consists of naturally occurring examples from spoken and written corpora in the two languages: two small corpora (BNC-Baby, CORLEC, CESJD-UCM), and two large corpora (BNC, CREA). In order to ascertain a basic level of *tertium comparationis* (Chesterman 1998), or of maximum similarity, at the level of the corpora, the study applies comparison criteria for the selection of the texts on the basis of relevant similarity constraints or factors (e.g. mode, genre and text type, and subject matter or topic in the CESJD-JMA corpus, though the level of expertise might differ in the oral unscripted corpora) that might affect the expression of evidentiality. In order to strive for the *tertium comparationis* at the semantic or notional level, the basic set of core evidential expressions (verbs and sentence adverbs) will be examined for cross-linguistic correspondence across the two languages, English to Spanish and Spanish to English, and the two paradigms will form the potential cross-linguistic paradigm of corresponding evidential expressions (Chesterman 1998; Altenberg & Granger 2002).

The study on multifunctionality is carried out in two stages:

(i) Pilot study with data from two corpora of unscripted conversation discourse in English and in Spanish, and an adhoc corpus of journalistic discourse in the two languages: (a) Oral: *BNC-Baby*-Unscripted conversation & *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC, UAM) (Mostly unscripted conversation); (b) Written: *Corpus of English and Spanish Journalistic Discourse* (CESJD-JMA) (comparable corpus of journalistic texts: opinion columns, leading articles, and news-reports).
(ii) Case study with data from the following corpora: (a) Oral: *BNC-World Edition* (Oral subcorpus) & *Corpus de Referencia del Español Actual* (CREA, RAE) (Oral subcorpus: Spain); (b) Written: *BNC-World Edition* (Newspaper subcorpus) & *Corpus de referencia del español actual* (CREA, RAE) (Newspaper subcorpus: Spain).

The analytical protocol for the annotation procedure draws on well-established criteria in the literature on evidentiality (cf. Anderson 1986; Diewald & Smirnova 2010; Boye 2012). The process of analysis of the data also ensures maximum comparability and involves: (i) manual, textual-based annotation of the selected evidential expressions, identifying the values of evidentiality (direct, inferential and reportative functions), quantification (using Monoconc), and statistical analysis of the quantitative results in the use of these resources across discourse domains (oral vs written discourse) and languages (English vs. Spanish); (ii) identification of those evidential expressions exhibiting multifunctionality and comparison of the quantitative results in relation to discourse domains and genres; (iii) identification of evidential values characteristic of the various construction types with verbal evidential expressions, and comparison of the quantitative results in both languages.

Preliminary results point to similarities across languages in the pattern of distribution and frequencies of evidential expressions. There are however significant differences in the presence and frequencies of evidential values between discourse-domains and journalistic genres in both languages. Regarding multifunctionality, certain distinctions have been observed in cross-linguistic terms. There are also distinctions in the correlations between evidential values and construction types.

It will be argued that the use of particular values of evidentiality in the discourse is sensitive to variation in discourse domains and genres, and that the existence of most salient values for particular evidential value-construction pairings would appear to indicate a process of entrenchment. Reference will also be made to certain parameters which may play a crucial role in facilitating these extensions of values within the domain of evidentiality, from inferential to reportative, resulting in the multifunctionality of particular evidential expressions (Marín-Arrese 2017).

References

Aikhenvald, A. (2004). *Evidentiality*. Oxford: Oxford University Press.

Altenberg, B. & Granger, S. (2002). Recent trends in cross-linguistic lexical studies. In B. Altenberg & S. Granger (eds). *Lexis in Contrast*. Amsterdam: John Benjamins, 3-48.

Anderson, L. (1986). Evidentials, paths of change, and mental maps: Typologically regular asymmetries. In W. Chafe & J. Nichols (ed.) *Evidentiality: The Linguistic Coding of Epistemology*. Norwood, NJ: Ablex, 273-312.

Boye, K. (2012). *Epistemic meaning: A crosslinguistic and functional-cognitive study*. Berlin: Mouton de Gruyter.

Chafe, W. (1986). Evidentiality in English conversation and academic writing. In W. Chafe & J. Nichols (eds). *Evidentiality: The Linguistic Coding of Epistemology*. Norwood, NJ: Ablex, 261-272.

Chafe, W. & J. Nichols (eds). (1986). *Evidentiality: The Linguistic Coding of Epistemology*. Norwood, NJ: Ablex.

Chesterman, A. (1998). *Contrastive Functional Analysis*. Amsterdam: John Benjamins.

Diewald, G. & Smirnova, E. (2010). *Evidentiality in German. Linguistic Realization and Regularities in Grammaticalization*. Berlin: Mouton de Gruyter.

Fitneva, S. (2001). Epistemic marking and reliability judgements: Evidence from Bulgarian. *Journal of Pragmatics* 33, 401-420.

Langacker, R. W. (2017). Evidentiality in Cognitive Grammar. In J. I. Marín-Arrese, G. Hassler & M. Carretero (eds). *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-Pragmatic Perspectives*. Amsterdam & Philadelphia: John Benjamins, 13-55.

Marín-Arrese, J. I. (2011). Epistemic Legitimising Strategies, Commitment and Accountability in Discourse. *Discourse Studies* 13 (6), 789-797.

Marín-Arrese, J. I. (2017). Multifunctionality of evidential expressions in discourse domains and genres: Evidence from cross-linguistic case studies. In J. I. Marín-Arrese, G. Hassler & M. Carretero (eds). *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-Pragmatic Perspectives*. Amsterdam & Philadelphia: John Benjamins, 195-223.

Papafragou, A., Li, P., Choi, Y. & Han, Ch. (2007). Evidentiality in Language and Cognition. *Cognition* 103(2), 253-299.

Key clusters as potential indicators of translator style

Lorenzo Mastropierro
University of Nottingham
lorenzo.mastropierro@nottingham.ac.uk

Translation has traditionally been seen as a derivative process, an act of reproduction, an inferior copy of the original. Many theorists in translation studies have overturned this view, advocating greater visibility of the translator in the translated text (Venuti 2008). Far from simply being a derivative reproduction, translation involves the active and creative contribution of the translator. The translator "rewrites" (Lefevere 1992) the source text, and this rewriting is never a passive transfer; rather, it is a form of (re)interpretation that involves both the linguistic and extra-linguistic level of the original. When we read a translated text, we do not engage with the original author's voice only, but also with that of the translator (Hermans 1996). Their presence can be overtly discernible, as in the case of paratextual commentaries, or, more often, hidden behind that of the author. This makes spotting the style of the translator challenging, as it is likely to be entangled to that of the original writer (Bernardini 2005). Yet, studying translator style is not only possible but also necessary: as Baker (2000: 262) explains, if we want to claim convincingly that translation is a creative enterprise, we need to explore more the question of style from the translator's point of view. This paper reports on a study that uses corpus methods to analyse translator style. It focuses on disambiguating and analysing linguistic features of translated texts that can be seen as independent from the original, and can therefore be attributed to the translators.

Previous corpus-based studies on translator style have compared the use of manually selected linguistic features across different target texts (see, for example, Baker 2000; Saldanha 2011; Winters 2010; Li et al. 2011). Other studies (Winters 2009; Wang & Li 2012) have instead used a key word analysis to obtain a computer-generated list of items that characterise one translation compared to the other. This paper builds on these studies but, differently from them, proposes a method to identify potential indicators of translator style based on key cluster analysis. 'Clusters' are sequences of words that are found repeatedly together (Scott 2016), such as *as if*, a 2-word cluster, *the middle of*, a 3-word cluster, *as I said before*, a 4-word cluster, etc. A key cluster analysis uses cluster frequency lists, instead of word lists, as starting point for the comparison. The computer compares the frequency of the clusters occurring in each text and identifies which clusters are used significantly more frequently in one translation compared to the other. The picture of the divergences between target texts that key clusters offer is not limited to the preference for the individual word, but can also encompass more complex lexico-grammatical structures. Multi-word sequences can in fact have a larger semantic and grammatical impact on meaning than individual words (Stubbs 2002, 2007) and their analysis has represented a major methodological approach to the study of stylistic relevance (Biber 2011: 17). To test the extent to which key clusters can indicate potential stylistic differences between translations of the same source text, I compare two different Italian translations of H. P. Lovecraft's *At the Mountains of Madness*. The first is the translation by Giuseppe Lippi, published by Mondadori in 1992; the second is the translation by Serenella Antonucci, published in 1993 by Newton & Compton. Comparing two versions of the same original allows me to bypass what the two translations share – which is likely to be related to the content and style of the source text – and focus instead on what differentiates the two texts.

The resulting key clusters (Table 1) are analysed as potential indicators of the translators' style. In particular, this paper selects and investigates key clusters that can be grouped together into related categories, as categories can signal – more efficiently than individual items – the presence of a pattern in the use of a given linguistic feature by a translator.

Three categories are identified, related to the use of Italian euphonic -*d* (for example *ad un*, “to a”; *ed il*, “and the”; *ed in*, “and in”), locative clitics (*c’è*/*vi è*, “there is”; *c’erano*/*vi erano*, “there were”), and distal demonstratives (for example *in quel*, “in that”, *in quel momento*, “in that moment”; *di quella*, “of that”).

Antonucci vs. Lippi				Lippi vs. Antonucci			
Key cluster	Freq.	RC. Freq.	Log-L	Key cluster	Freq.	RC. Freq.	Log-L
ad un	38	0	51,50	a un	41	6	30,36
ad una	27	0	36,60	e io	19	0	26,94
ed il	25	0	33,88	a una	30	3	26,49
ed i	25	0	33,88	d anni	16	0	22,68
in quel	64	16	29,36	fra le	21	2	18,89
in quel momento	30	3	24,81	e altri	13	0	18,43
così come	16	0	21,69	un attimo	13	0	18,43
al di	62	20	21,28	danforth e io	13	0	18,43
di quella	33	6	19,75	la creatura	12	0	17,01
per cui	37	8	19,37	all accampamento	16	1	16,43
di quelle	39	10	17,45	e il	47	17	15,58
vi erano	17	1	16,73	c era	25	5	15,19
ed in	12	0	16,26	alle spalle	15	1	15,14
un istante	12	0	16,26	c erano	15	1	15,14
quel momento	50	17	15,97				
di quel	35	9	15,61				

Table 1. Key clusters

The contrastive analysis of the clusters reveals some stylistic idiosyncrasies of the translators. These idiosyncrasies are analysed in terms of the stylistic effect they convey, highlighting the potential impact they can have on the translated text from the Italian reader’s point of view. In particular, it is shown that the larger use of euphonic -*d* and *vi* locative clitics contribute to give Antonucci’s translation a more dated and formal feeling, whereas the more frequent occurrence of distal demonstratives has an effect on the text deixis and on the representation of the narrator’s point of view.

References

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

Bernardini, S. (2005). Reviving old ideas: Parallel and comparable analysis in translation studies – with an example for translation stylistics. In K. Aijmer & C. Alvstad (eds). *New Tendencies in Translation Studies*. Göteborg: University of Göteborg, 5-18.

Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature* 1(1), 15-23.

Hermans, T. (1996). The translator’s voice in translated narrative. *Target* 8(1), 23-48.

Lefevere, A. (1992). *Translation, Rewriting, and the Manipulation of Literary Fame*. London: Routledge.

Li, D., Zhang, C. & Liu, K. (2011). Translation style and ideology: A corpus-assisted analysis of two English translations of Hongloumeng. *Literary and Linguistic Computing* 26(2), 153-166.

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

Scott, M. (2016). WordSmith Tools Help. Liverpool, UK: Lexical Analysis Software. Accessed January 05, 2017. http://www.lexically.net/downloads/version6/HTML/
index.html?getting_started.htm.

Stubbs, M. (2002). *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stubbs, M. (2007). Quantitative data on multi-word sequences in English: The case of the word *world*. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (eds). *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum, 163-189.

Venuti, L. (2008). *The Translator’s Invisibility: A History of Translation (2nd ed.)*. London: Routledge.

Wang, Q. & Li, D. (2012). Looking for translator’s fingerprints: A corpus-based study on Chinese translation of *Ulysses*. *Literary and Linguistic Computing* 27(1), 81-93.

Winters, M. (2009). Modal particles explained: How modal particles creep into translations and reveal translator’s styles. *Target* 21(1), 74-97.

Winters, M. (2010). From modal particles to point of view: A theoretical framework for the analysis of translator attitude. *Translation and Interpreting Studies* 5(2), 163-185.

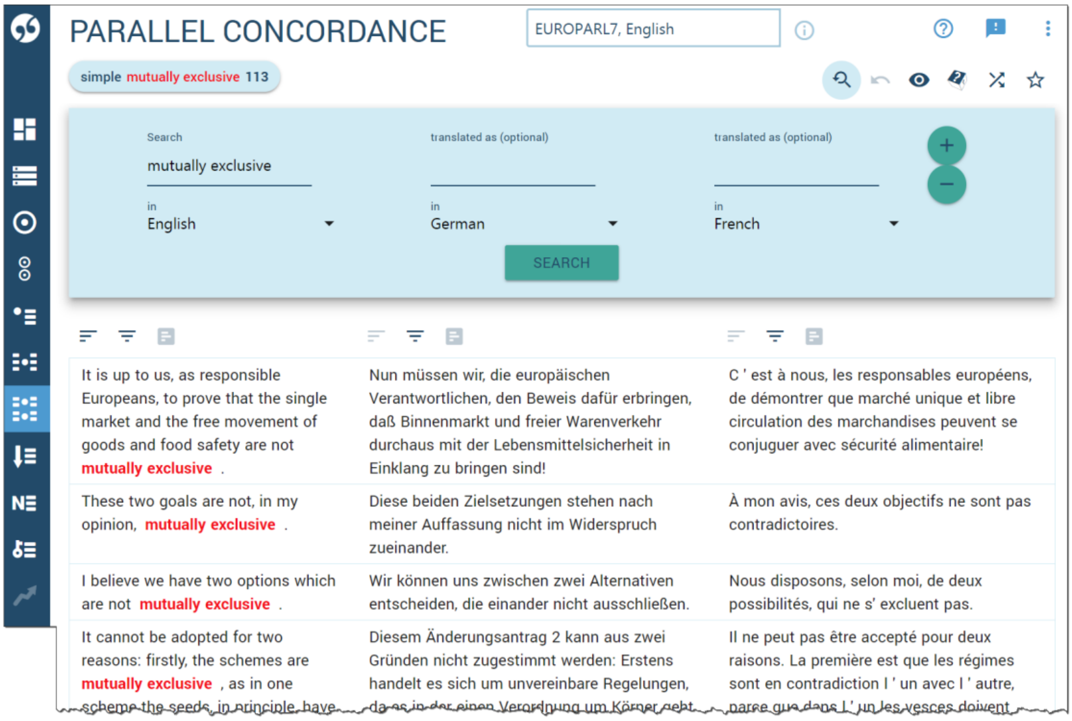
Bilingual functionalities of Sketch Engine

Ondřej Matuška
Lexical Computing
ondrej.matuska@sketchengine.eu

Sketch Engine (a product of Lexical Computing) is a corpus management and corpus query software which houses a suite of over 400 corpora in 95+ languages and supports 20+ writing systems. Users can take advantage of the built-in automated corpus building tool which is optimized for users with little or no technical knowledge of corpus building, tagging or lemmatizing.

Sketch Engine started as a specialized lexicographic tool but over time it has developed into a tool that can be used by anyone who needs to understand how language is used. With the addition of new features, the user base extended to include linguists and other researchers from a variety of fields such as social sciences and humanities. The development of the new interface opened the doors for new groups of users including translators, copywriters, marketing specialists, authors or teaching materials, teachers or learners of languages (Kilgarriff et al. 2014).

What makes Sketch Engine stand out is its support of multiple languages including parallel multilingual texts. While acquiring parallel data in the past was nearly impossible or involved a tedious manual process, the situation changed now with institutions releasing their translated documents electronically, often in aligned formats. Aligned multilingual texts are now daily produced by translators in their Computer Assisted Translation (CAT) tools. Such data are ready to be converted into a multilingual corpus ready for contrastive studies. One way of looking at multilingual data is parallel concordance, cf. Figure 1 where the search results are displayed side by side. The result screen shows how the English phrase was translated into German and French. Search criteria can be set for each language independently. The results can be processed further by sorting, filtering or calculating frequencies. While the typical use assumes two or three languages to be displayed at the same time, Sketch Engine supports parallel corpora with an unlimited number of aligned languages and all of them can be searched at the same time and displayed next to each other.



Search: mutually exclusive
in: English

translated as (optional):
in: German

translated as (optional):
in: French

It is up to us, as responsible Europeans, to prove that the single market and the free movement of goods and food safety are not **mutually exclusive** .	Nun müssen wir, die europäischen Verantwortlichen, den Beweis dafür erbringen, daß Binnenmarkt und freier Warenverkehr durchaus mit der Lebensmittelsicherheit in Einklang zu bringen sind!	C ' est à nous, les responsables européens, de démontrer que marché unique et libre circulation des marchandises peuvent se conjuguer avec sécurité alimentaire!
These two goals are not, in my opinion, **mutually exclusive** .	Diese beiden Zielsetzungen stehen nach meiner Auffassung nicht im Widerspruch zueinander.	À mon avis, ces deux objectifs ne sont pas contradictoires.
I believe we have two options which are not **mutually exclusive** .	Wir können uns zwischen zwei Alternativen entscheiden, die einander nicht ausschließen.	Nous disposons, selon moi, de deux possibilités, qui ne s' excluent pas.
It cannot be adopted for two reasons: firstly, the schemes are **mutually exclusive** , as in one scheme the seeds, in principle, have	Diesem Änderungsantrag 2 kann aus zwei Gründen nicht zugestimmt werden: Erstens handelt es sich um unvereinbare Regelungen, da es in der einen Verordnung um Körner geht	Il ne peut pas être accepté pour deux raisons. La première est que les régimes sont en contradiction l ' un avec l ' autre, parce que dans l ' un les vesces doivent

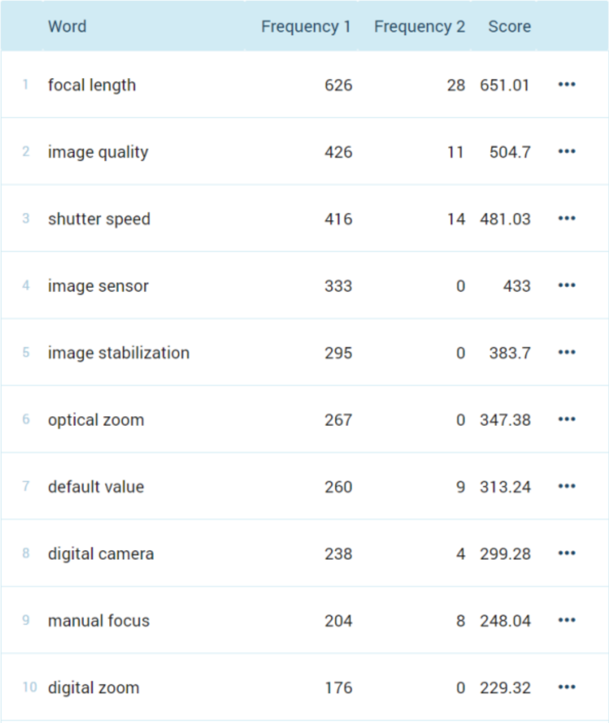
Figure 1. Result screen for mutually exclusive

In addition to processing texts using statistics, Sketch Engine also uses linguistic criteria such as parts of speech and grammatical categories. This is facilitated by part-of-speech tagging that is carried out automatically on each uploaded text. In addition, automatic lemmatization simplifies the work with texts in morphologically rich languages. The combination of statistics with linguistic criteria gave rise to some of the best term extraction technologies, offering also the bilingual term extraction option (Kovář et al. 2016).

The traditional methods of terminology extraction rely on identifying high-frequency multi-word expression. This leads inevitably to lists polluted by frequently used expressions which are not terms and a thorough manual cleaning process is necessary.

Sketch Engine starts with linguistic criteria first. There are language specific rules, called *term grammar*, for each language where term extraction is supported. The term grammar sets the permitted format of a term in the language. The English term grammar will contain rules such as noun+noun, noun+*of*+noun, adjective+noun, adjective+noun+noun etc. The Spanish term grammar will contain noun+adjective, noun+*de*+noun, noun+*de*+noun+adjective, etc. These examples are only illustrative and the actual term grammar is noticeably more complex.

Sketch Engine will first identify all lexical units that comply with the term grammar before moving on to frequencies. In the next step, the frequency of each such lexical unit is computed in the target text. Then, Sketch Engine will calculate the frequency of the same lexical unit in general language. Large multi-billion-word corpora are used as samples of general language (reference corpora) because they cover such a wide range of texts and topics as to be considered representative of language in general. Frequencies in the target text are compared to the frequencies in the reference corpus and lexical units whose frequency is higher in the target text than it is in language generally are presented to the user as terms. This approach produces a very clean list of terms which requires very little manual cleaning (Figure 2).



	Word	Frequency 1	Frequency 2	Score	
1	focal length	626	28	651.01	•••
2	image quality	426	11	504.7	•••
3	shutter speed	416	14	481.03	•••
4	image sensor	333	0	433	•••
5	image stabilization	295	0	383.7	•••
6	optical zoom	267	0	347.38	•••
7	default value	260	9	313.24	•••
8	digital camera	238	4	299.28	•••
9	manual focus	204	8	248.04	•••
10	digital zoom	176	0	229.32	•••

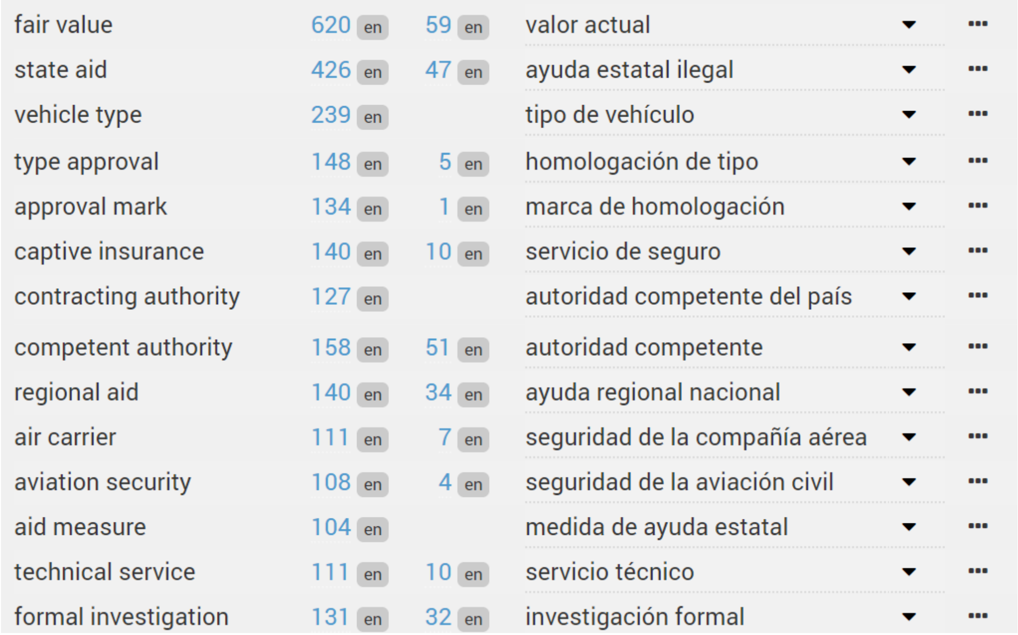
Figure 2. Output of the term extraction functionality

Figure 2 shows the output as it comes out of the system without any manual cleaning applied. Terms were extracted from texts about digital cameras. The typical user of the term extraction functionality is a practising translator for whom the choice of settings within the Sketch Engine interface might be overwhelming. This is true especially as the default settings already produce a high-quality output. A decision was made to make the term extraction more user-friendly and rework it into a one-step procedure (Jakubíček et al. 2014).

A completely separate term extraction interface was developed and called OneClick Terms. The user only needs to select a file (or drag&drop it into the interface) and click a button. The term extraction will start and the process does not require any additional user intervention for the terms to be extracted. When translation memory (tmx) is uploaded, a bilingual term extraction option is offered.

Figure 3 shows terms and their translations. The user can select a different translation. The local menu gives access to the relevant Wikipedia pages and also the terms in context in both languages.

Results can be downloaded for import into other software such as spreadsheet, CAT tool or term management tool.



fair value	620	en	59	en	valor actual	▼	•••
state aid	426	en	47	en	ayuda estatal ilegal	▼	•••
vehicle type	239	en			tipo de vehículo	▼	•••
type approval	148	en	5	en	homologación de tipo	▼	•••
approval mark	134	en	1	en	marca de homologación	▼	•••
captive insurance	140	en	10	en	servicio de seguro	▼	•••
contracting authority	127	en			autoridad competente del país	▼	•••
competent authority	158	en	51	en	autoridad competente	▼	•••
regional aid	140	en	34	en	ayuda regional nacional	▼	•••
air carrier	111	en	7	en	seguridad de la compañía aérea	▼	•••
aviation security	108	en	4	en	seguridad de la aviación civil	▼	•••
aid measure	104	en			medida de ayuda estatal	▼	•••
technical service	111	en	10	en	servicio técnico	▼	•••
formal investigation	131	en	32	en	investigación formal	▼	•••

Figure 3. Bilingual term extraction output

References

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2014). Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 53-56.

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

Kovář, V., Baisa, V. & Jakubíček, M. (2016). Sketch Engine for bilingual lexicography. *International Journal of Lexicography* 29(3), 339-352.

Parallel corpus vs. bilingual dictionary: Their usefulness in translator training

Adriana Mezeg
University of Ljubljana
adriana.mezeg@ff.uni-lj.si

In translator training, we allow students to use all kinds of language tools and resources in order to facilitate their translation tasks, for example monolingual, bilingual or multilingual dictionaries and glossaries, monolingual (national), comparable and parallel corpora, parallel and comparable texts, translation memories, terminological databases, forums, etc. However, according to our translator training practice, students nowadays still mostly use electronic dictionaries which they often combine with hybrid resources such as Linguee and with corpora. Excluding the Linguee web service because its construction is not controlled in a way that of dictionaries and corpora is and because the quality of its translations is uncertain, the present paper would like to examine how useful a parallel corpus is in comparison with a bilingual dictionary in concrete translation tasks, and identify their main (dis)advantages.

The study is based on 10 different texts (each containing from 400 to 1000 words) translated by 18 MA students in the academic year 2016/2017 as part of their school or home assignments in the framework of the seminar Translation of specialised texts from French into Slovenian at the Department of Translation, Faculty of Arts, University of Ljubljana. The seminar covers different types of texts (e.g. newspaper articles, contracts, scientific articles and reports, EU documents, financial documents, instruction manuals) containing vocabulary specific to the fields of politics, economics, law, mechanical engineering, medicine, etc., but also general words and expressions that can be found in any text. On the one hand, we analysed translations of individual texts students had to make throughout the academic year; on the other hand, we examined their seminar papers consisting of a translation of a chosen source text, a thorough description of the translation process and an explanation of their translation solutions supported by the resources used. The ten papers appear in the book *Kaj se skriva za prevodom? Izzivi pri prevajanju iz francoščine v slovenščino – zbornik študentskih prevajalskih nalog* (*What is there behind a translation? Challenges in translating from French into Slovenian – a book of students' translation projects*), published in Slovenian in 2017.

In this study, we wish to focus on the utility of a concrete parallel corpus and bilingual dictionary, to date the only or the most comprehensive resources for the language pair French-Slovenian, i.e. a) the French-Slovenian parallel corpus FraSloK (around 2.5 million words), containing, on the one hand, 300 articles from *Le Monde diplomatique* (637,297 words) and their Slovenian translations (526,777 words) and, on the other hand, 12 contemporary French novels (701,715 words) along with their Slovenian translations (601,196 words) (we are aware that the literary part can mostly be useful for general vocabulary, whereas the journalistic part may serve us when we are looking for a more specific lexis, particularly that from the fields covered by *Le Monde diplomatique*, for example politics, economy, law), and b) the digitised French-Slovenian dictionary (42,000 entries and subentries) by Anton Grad, first published in 1975 and to date still the largest dictionary for the language pair in question, which has not yet been updated or replaced by a modern one of similar or larger size, mostly for financial reasons. Despite its year of publication, the dictionary is, generally speaking, a decent starting point when we want to verify the meaning or possible translation equivalents of an unknown French word or expression.

With regard to some problematic or difficult individual words or expressions used in the French texts, some of which will be discussed in the paper, we compared the utility of the mentioned resources. The "problematic" vocabulary, i.e. the French words or expressions students had trouble translating into Slovenian, were mostly from the fields of politics (e.g. *régime*, *administration*, *porte-parole*, *état-major*, *cessez-le-feu*, *fraude*) and economics (*précarité*, *service*, *prestataire*, *commercialiser*, *support de diffusion*), but it also concerned some general lexis (e.g. *sondage*, *d'urgence*, *couverture médiatique*, *front*, *motivation religieuse*, *forfait*, *volet*) and that from the field of medicine (e.g. *excision*).

Overall, as far as the translations made in class or at home go, the Grad dictionary appeared to be less useful than the parallel corpus, mostly for the following reasons: a) because it did not contain a word searched for (we only examined words that started to be used in French before 1975 and could thus be included in the Grad French-Slovenian dictionary); b) because sometimes none of the translation equivalents corresponded to what the students were looking for; c) because the proposed translation equivalents were awkward, literal translations, not in line with a word or expression a native Slovenian speaker would actually use. In this respect, the FraSloK corpus turned out to be more helpful as it included translation equivalents for most of the words inquired, be it general or specialised, but mostly because in a corpus, the words or expressions are surrounded with co-text and so their meaning and use are easier to detect. However, sometimes a searched word or expression was not found in either of the resources used, so students had to find other ways in quest of a translation.

In conclusion, while the bilingual dictionary turned out to be less useful for lack of some French entries and/or the context (no examples of use provided), particularly when several translation equivalents were listed for a searched word, the usefulness of a parallel corpus, according to this study, depended to a large extent on its size and the variety of text types included (a corpus of about 1.3 million French words used in journalistic and literary texts might not include a word or expression that is rarely used or used only in specific contexts, etc. (e.g. *une trémie* (*a hopper*), *la noosphère* (*the noosphere*), *ventiler* (*to break down*)). The papers from the book of students' translation projects confirm this and show that the available French-Slovenian dictionary and the FraSloK corpus do not meet all translation needs, which is undoubtedly true. In fact, it is unlikely we will ever get such an almighty language tool/resource that would meet all translator's needs As for the language pair French-Slovenian, we should first at least revise and enlarge the current dictionary and add new text types to the existing parallel corpus. Meanwhile, according to the questionnaires our students had to fill in at the end of the academic year in relation to the texts translated during the winter and spring semester, the following might be of help: a good monolingual dictionary, a comprehensive French-English dictionary as a good starting point to find an English equivalent and then continue the search, for example, in an English-Slovenian dictionary (for example in the *Veliki angleško-slovenski slovar Oxford* edited by Simon Krek in 2005, which is modern, comprehensive and often includes examples of use) to increase the possibilities of finding the adequate Slovenian translation equivalent, parallel and comparable texts, consultation with a specialist in a certain field and, last but not least, general culture, common sense and logical thinking.

References

Grad, A. (1975). *Francosko-slovenski slovar*. Ljubljana: DZS.

Jesenik, V. & Dembskij, N. (1990). *Slovensko-francoski slovar*. Ljubljana: DZS.

Krek, S. (ed.) (2005). *Veliki angleško-slovenski slovar Oxford*. Ljubljana: DZS.

Mezeg, A. (2010). Compiling and Using a French-Slovenian Parallel Corpus. In R. Xiao (ed.) *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2010)*. Ormskirk: Edge Hill University, 1-27. http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2010 Proceedings/papers/Mezeg.pdf (Accessed 28/11/2017).

Mezeg, A. (ed.) (2017). *Kaj se skriva za prevodom? Izzivi pri prevajanju iz francoščine v slovenščino – zbornik študentskih prevajalskih nalog*. Ljubljana: Znanstvena založba Filozofske fakultete.

Mikhail, M. & Cooper, R. (2016). *Corpus linguistics for translation and contrastive studies: a guide for research*. Abingdon & New York: Routledge.

Pisanski, A. (2003). Uporaba novih tehnologij pri jezikovnem pouku. *Jezik in slovstvo* 48(3-4), 103-112.

Vintar, Š. (2001). Računalniška orodja za jezikoslovce in prevajalce. *Zbornik 37. Seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta, 319-332.

Vintar, Š. (2008). Corpora in Translation: A Slovene Perspective. In G. M. Anderman & M. Rogers (eds). *Incorporating corpora: the Linguist and the Translator*. Clevedon, Buffalo & Toronto: Multilingual Matters, 153-167.

Explicitation revisited: Shifts in the translation of reporting verbs in the French-English-Czech part of the InterCorp parallel corpus

Olga Nádvorníková
Charles University
olga.nadvornikova@ff.cuni.cz

The usage of reporting verbs in fiction varies considerably from language to language (*cf.* Tegelberg 1999 for French/Danish; Fónagy 1986 for English/Hungarian; Corness 2010 and Pípalová 2012 for English/Czech; Fárová 2016 for English/Czech/Finnish or Nádvorníková 2017a for French/English/Czech). In English reporting clauses, the neutral *he said* is preferred to more specific forms such as *he grumbled, gasped, cautioned* or *lied*. In Czech, on the contrary, the repetition of the reporting verb *říci* (impf. *říkat*), equivalent of *say*, is considered clumsy (Levý & Jettmarová 2011). Subsequently, authors (and translators) try to avoid it, using not only synonyms of *say/říci*, but also other verbs, explicitating various aspects of the reported speech (*odpovědět/answer*, *úpět/groan,* etc.) (see examples 1a-c).

(1a) "My wand," **said** Ron, in a shaky voice. (J.K. Rowling, *Harry Potter and the Chamber of Secrets*, 1998)
(1b) *– Ma baguette, **répondit [answered]** Ron d'une voix tremblante.* (transl. by J.F. Ménard, 1999)
(1c) *"Moje hůlka," **úpěl [groaned]** Ron roztřeseným hlasem.* (transl. by P. Medek, 2000)

A recent study carried out on the InterCorp parallel corpus (Nádvorníková 2017a) has shown that the frequency of explicitation of a neutral reporting verb (*say/dire/říci*) is higher in translations from languages with a higher proportion of neutral verbs in reporting clauses (English 60%, French 50%) into a language where their proportion is lower (Czech 30%), than in the opposite direction of translation. This result clearly shows that in this case, explicitation is a language-pair dependent phenomenon (cf. House 2008), and not a universal feature of translation (Blum-Kulka 1986; Baker 1993; Øverås 1998, etc.). This observation is corroborated also by a complementary search in the Czech comparable translation corpus Jerome (Chlumská 2013), showing that Czech translated texts respect the same stylistic norm as the non-translated ones, even at the cost of explicitation or translation shifts (indeed, if they want to achieve the proportion of neutral reporting verbs usual in the target language, Czech translators have to use all means of variation of reporting verbs). Nevertheless, this basic study has left aside several important issues that I will try to treat in this paper:

1. a more thorough **quantitative analysis** of lexical variation of reporting verbs is needed, as the simple proportion of neutral reporting verbs *say/dire/říci-říkat* is not sufficient. We assume that a high proportion of the neutral reporting verb will correlate with a low type/token ratio of all reporting verbs;
2. the analysis of explicitation cannot be complete without the analysis of its counterpart, **implicitation**: in accordance with the asymmetry hypothesis (Klaudy & Károly 2005), we assume that the frequency of implicitation in one translation direction will be lower than the frequency of explicitation in the opposite translation direction;
3. more importantly, a more fine-grained **typology of explicitations** (changes) occurring in translations of reporting verbs is necessary (cf. Klaudy & Károly 2005, the concept of "intervention" in House 2008, etc.).

We assume that systemic and typological differences between the three languages may play an important role, e.g. the disappearance of transgressive (Czech adverbial non-finite verb form) from the Czech language system (Nádvorníková 2010), causing changes of "information packaging" in Czech (cf. also Nádvorníková 2017b) (see examples 2a-c).

(2a) 'The birds again!' **said** Aragorn, **pointing down.** J.R.R Tolkien, *The Fellowship of the Ring*, 1954)
(2b) *– Les oiseaux, encore ! **dit** Aragorn, **pointant le doigt**.* (transl. by F. Ledoux, 1988)
(2c) *"Zase ti ptáci!" **ukázal dolů [pointed down]** Aragorn.* (transl. by S. Pošustová, 1990)

In our typology of explicitations, we will concentrate on the syntactic and semantic changes (especially raising of participial phrases to clause level and semantic specification of reporting verbs) in combination with the (difficult) distinction between the obligatory and the optional shifts.

The InterCorp parallel corpus (www.korpus.cz/intercorp) is used as source of data in this research: first as unidirectional – for the quantitative analysis of lexical variation of reporting verbs in the three languages (*t/t ratio*, etc.), then as bi/multidirectional – for the translation counterparts, necessary in the research of implicitation and the typology of explicitation. The monolingual parts of InterCorp, even when limited to the non-translated fiction, contain 18,953,496 corpus positions (words and punctuation marks) for English, 7,161,043 positions for French and 18,627,345 positions for Czech. The intersections of the three languages are of course less representative, but sufficient as well (they contain in average 5,824,860 corpus positions; the most for translations from English to Czech, 18,953,496 positions, the fewest for translations from English to French, 573,088 positions). For each direction of translation (*en-fr, fr-en, cs-fr, fr-en, en-cs, cs-en*), we analyzed manually 1,000 occurrences of reporting verbs in medial and postposed reporting clauses, i.e. 6,000 occurrences in total. We observed not only the translation counterparts of reporting verbs, but also the use of adjuncts, changes of the position of the reporting clause, etc.

We are aware of the fact that the translation of reporting verbs may be influenced not only by the linguistic and stylistic factors, mentioned above, but also by a specific translation tradition, editorial changes, prestige of the source/target text literature or by authors'/translators' idiolects. Nevertheless, we hope that a thorough quantitative and qualitative analysis of the translation of reporting verbs will shed more light on the causes and consequences of explicitation in translation in general.

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: John Benjamins, 233-250.

Blum-Kulka, S. (1986). Shifts of Cohesion and Coherence in Translation. In J. House & S. Blum-Kulka (eds). *Interlingual and Intercultural Communication: discourse and cognition in translation and second language acquisition studies*. Tübingen: Gunter Narr, 17-35.

Chlumská, L. (2013). *JEROME: srovnatelný korpus překladové a nepřekladové češtiny*. Praha: ÚČNK FF UK, http://www.korpus.cz.

Corness, P. (2010). Shifts in Czech translations of the reporting verb said in English fiction. In F. Čermák, P. Corness & A. Klégr (eds). *InterCorp: Exploring a Multilingual Corpus*. Praha: Nakladatelství Lidové noviny, 159-177.

Fónagy, I. (1986). Reported speech in French and Hungarian. In F. Coulmas (ed.) *Direct and Indirect Speech*. Berlin: Mouton de Gruyter, 255-311.

House, J. (2008). Beyond Intervention: Universals in Translation? *trans-kom*, 1, 6-19.

Klaudy, K. & Károly, K. (2005). Implicitation in Translation: Empirical Evidence for Operational Asymmetry in Translation. *Across Languages and Cultures*, 6(1), 13-28.

Levý, J. & Jettmarová, Z. (2011). *The art of translation*. Amsterdam & Philadelphia: John Benjamins.

Nádvorníková, O. (2010). The French gérondif and its Czech Equivalents. In Fr. Čermák, P. Corness & A. Klégr (eds). *InterCorp: exploring a multilingual corpus*. Praha: Nakladatelství Lidové noviny, 83-95.

Nádvorníková, O. (2017a). Les proportions des verbes SAY/DIRE/ŘÍCI dans les propositions incises et leurs équivalents en traduction : étude sur corpus parallèle. *Linguistica Pragensia* 28(2), 35-57.

Nádvorníková, O. (2017b). Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus. In J. Emonds & M. Janebová (eds). Olomouc & Palacký University, 445-461. Available at http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016- proceedings.pdf.

Øverås, L. (1998). In Search of the Third Code: An Investigation of Norms in Literary Translation. *Meta: Tranlator´s Journal* 43(4), 557-570.

Pípalová, R. (2012). Framing Direct Speech: Reporting Clauses in a Contrastive Study. *Prague Journal Of English Studies* 1(1), 75-107.

Tegelberg, E. (1999). Les verbes d'incise dans Hemsöborna et sa traduction française. Étude contrastive. *Studia Neophilologica* 71, 72–96.

Hypal4MUST: An enhanced user-friendly tool for the collection, alignment and annotation of student translations

Adam Obrusnik[1], Marie-Aude Lefer[2]
Hypal/Hypal4MUST developer[1], Université catholique de Louvain[2]
admin@hypal.eu, marie-aude.lefer@uclouvain.be

Learner translation corpora (LTC), i.e. parallel corpora containing translations produced by students, are highly valuable teaching and research resources, but they require demanding text pre-processing. The main reason for this is that LTC are so-called *multiple* translation corpora, i.e. corpora where each source text corresponds to several translations. The present software demo aims at introducing Hypal4MUST, a web-based user-friendly tool for the collection, alignment and linguistic annotation of LTC. Hypal4MUST is based on the Hypal[1] tool (Fictumova et al. 2017). It contains a broad range of additional functionalities tailor-made for the specific needs of the *Multilingual Student Translation* (MUST) project (Granger & Lefer 2017, this volume).

Hypal4MUST, like the original Hypal, introduces Web 2.0 technologies in LTC pre-processing. A distinctive feature of Hypal and Hypal4MUST, when compared to other software such as InterText (Vondřička 2014) and EasyAlign (Evert et al. 2016) for alignment, is its modular and responsive web-based interface. The interface integrates the pre-processing of LTC, i.e. the alignment, part-of-speech (POS) tagging and annotation of parallel texts, as well as the possibility of collecting translations and metadata directly from students (translation students and language learners). Each sub-corpus within Hypal and Hypal4MUST can act as a translation task for students, to which they can submit assignments via a separate simplified *student interface*. Assignments collected from students are then POS-tagged, aligned and annotated by teachers/researchers using the main interface (the so-called *teacher interface*). Teachers can share their feedback by returning annotated assignments to learners via the student interface.

From the technical standpoint, the software relies on an automatic text alignment algorithm which shares some of the features of the Gale-Church (Gale & Church 1993) and Hunalign (Varga et al. 2005) algorithms (e.g. it relies on normalized sentence length). The accuracy of automatic alignment can be improved by identifying lexical matches in the source and target texts (Obrusnik 2013). For this, a bilingual dictionary needs to be uploaded to Hypal4MUST. The output of automatic alignment at sentence level can be edited manually in the graphical interface. The annotation interface has been developed from scratch and allows translation-oriented text annotation at the level of individual tokens or sequences of tokens. Hypal4MUST currently supports POS-tagging for the following languages: Chinese, Czech, Dutch, English, French, German, Italian, Polish, Portuguese, Russian, Slovak and Spanish. The POS-tagging is optional for the alignment algorithm (it is necessary only if a bilingual dictionary is provided) but it substantially increases the value of the annotation statistics (e.g. being able to correlate an annotation tag with a specific lemma/POS, not only with a word form). For languages for which a POS-tagger is not available, Hypal4MUST can be used without POS-tagging. Rich-text formatting (such as text justification, different typefaces, headings, bulleted lists or tables) is supported throughout the whole process.

There are several new key functions of Hypal4MUST that reflect the needs of the MUST project and that are not, to the best of our knowledge, available in any other software. The first of these is the so-called *Source Text Database*. The Source Text Database is a collaborative repository of source texts for MUST partners. Project partners either submit their own source text or choose an existing one from the database when creating translation tasks for their students. Each text in the shared Source Text Database has been validated by MUST project directors, ensuring that the rich metadata supplied together with source texts is complete and accurate. Another practical feature of the Source Text Database is the possibility of uploading a file with reference translations (i.e. translations produced by professional translators or members of the teaching/research staff),

[1] The original Hypal software remains available for parallel text alignment and error annotation of both learner and parallel corpora.

which is then available to MUST partners together with the corresponding source texts. A rather unique feature of Hypal4MUST is its structured metadata forms. Since the metadata collected within the MUST project (source text-, translation task- and student-related metadata) is entered by several independent users (the teacher/researcher creating the translation task, the students submitting translations), Hypal4MUST is capable of generating forms for filling in the metadata dynamically, taking into account complex conditional relations between the metadata fields. For example, if the teacher allows the use of reference tools for a given translation task, the student is prompted to enter the tools used while translating (the internet, translation forums, monolingual and bilingual dictionaries, etc.). With regard to text annotation, Hypal4MUST includes a multi-level annotation system called *Translation-oriented Annotation System* (TAS) (e.g. ST-TT transfer > Content transfer > Distortion, Language > Lexis and terminology > Multiword non-term > Collocation) (see Granger & Lefer, this volume). TAS is currently structured so that its various annotation levels are applicable to all the languages represented in the MUST project, which enables cross-linguistic, language-independent comparisons of annotation statistics.

In our presentation, we will demo the teacher and student interfaces so as to illustrate the full workflow of Hypal4MUST.

References

Evert, S. & the CWB Development Team (2016). The IMS Open Corpus Workbench (CWB) Corpus Encoding Tutorial. CWB Version 3.4, http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial/.

Fictumova, J., Obrusnik, A. & Stepankova, K. (2017). Teaching Specialized Translation Error-tagged Translation Learner Corpora. *Sendebar* 28, 209-241.

Gale, W. A. & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1), 75-102.

Granger, S. & Lefer, M.-A. (2017). Bridging the gap between learner corpus research and translation studies: The Multilingual Student Translation corpus. Paper presented at the 4th *Learner Corpus Research Conference*, Bolzano, 5-7 October 2017.

Obrusnik, A. (2013). *A hybrid approach to parallel text alignment*. Bachelor Thesis. Masaryk University, Brno, Czech Republic.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, 590-596.

Vondřička, P. (2014). Aligning parallel texts with InterText. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Paris: European Language Resources Association (ELRA), 1875-1879.

Dialogue vs. narrative in fiction: A cross-linguistic comparison

Signe Oksefjell Ebeling, Jarle Ebeling
University of Oslo
s.o.ebeling@ilos.uio.no, jarle.ebeling@usit.uio.no

It has long been acknowledged that language use varies across registers, e.g. news, fiction, conversation, academic writing (see e.g. Biber et al. 1999; Stubbs & Barth 2003). In the context of contrastive analysis, cross-linguistic variation across registers has recently also received some attention (e.g. Lefer & Vogeleer 2014; Neumann 2014; Johansson 2007; Teich 2003), while variation *within* registers seems to have received less attention. This paper explores this somewhat neglected area by comparing features of dialogue and narrative in English vs. Norwegian fiction, thus adding a contrastive perspective to a question recently posed by Egbert & Mahlberg (2017): "Fiction – one register or two?"

In this paper, we use the term dialogue to mean passages marked by the writer as being instances of direct speech, and, to a much lesser extent, direct thought. Examples of dialogue would be the two clauses in quotation marks in (1):

(1) Sarah sat up straight.
"Let's hope it doesn't rain," she said.
"I don't mind a little rain," Macon said. (AT1)

The investigation draws on material from the English-Norwegian Parallel Corpus (ENPC) and exploits the fact that the original texts in the fiction part of the corpus have been marked up for dialogue. Thus, the proportion of tokens contained within dialogue vs. narrative can easily be established. The texts are shown to be predominantly narrative in nature in both languages. One of the main aims of the present study will therefore be to assess to what extent results from previous contrastive studies based on the ENPC, which typically do not distinguish between dialogue and narrative, can be said to be biased towards narrative, which by far represents the larger portion of tokens. And if so, in what way, and what implications may it have for the validity of previous studies?

Preliminary observations suggest that the three types of fiction contained in the ENPC – Children's Fiction, Detective Fiction and General Fiction – may behave differently with regard to the proportion of dialogue vs. narrative in the texts. To be able to draw on a more homogeneous material for this investigation, it was therefore decided to focus on a subset of the ENPC, namely the General Fiction part only.

The paper starts with some relevant overviews of the comparable data, including distribution of dialogue vs. narrative in the texts, followed by a comparison of language-specific n-word lists in dialogue vs. narrative. Results from these general surveys show that the Norwegian original texts generally contain less dialogue than the English ones. Moreover, the n-word lists suggest that dialogue is more repetitive than narrative in both languages, in the sense that the same types are used over and over again.

Two qualitative contrastive studies of items that have previously been investigated on the basis of the ENPC will be carried out in order to find out whether their use in dialogue differs from that in narrative and to what extent this comes out in translation. The items under scrutiny are *there* (Ebeling 2000) and *see* (Øhman 2006). The items were chosen precisely because they had been studied before, but also because both were found to be proportionally more frequent in dialogue.

Preliminary results from the two case studies uncover some differences in the use of *see* in dialogue vs. narrative. *See* is a highly polysemous verb (e.g. Alm-Arvius 1993; Aijmer 2004) and some meanings are more salient in dialogue than in narrative and vice versa. Notably, the "understand" meaning of *see*, as in *Do you see* *what I mean?*, is, not surprisingly, more frequently found in dialogue. This is clearly reflected in the translation paradigms of *see* in dialogue vs. narrative and will therefore have implications for cross-linguistic insights regarding the use of this lemma and its Norwegian correspondences. In the case of *there*, few differences in use can be detected on the basis of the material at hand. A similar proportion of existential vs. locative *there* is found in dialogue and narrative (approx. 68% ex. vs 28% loc. in dialogue and approx. 70% vs. 30% in narrative). This preference for the existential use is also reflected in the Norwegian correspondences, especially attested by the high number of congruent translations with existential *det*, or other ways of expressing existence. Other uses of *there*, e.g. phraseological ones or when the sequence *there, there* is used in order to comfort someone, as in *"There, there; good boy"*, are infrequent in both registers, but more commonly found in dialogue. This last point is interesting contrastively, since these more specialised uses of *there* seem to have standard correspondences typical of dialogue (*så, så* 'so, so' in the example above).

We take these observations to indicate that the need for splitting fiction into two registers – dialogue and narrative – may or may not be essential to gain more knowledge about the (cross-linguistic) characteristics of the language of fiction. In other words, depending on the object of study, fiction may or may not be seen to consist of two registers rather than one. It may therefore be advisable to keep this in mind before embarking on a contrastive study of fictional texts.

Although the present study shows that there may be a case for taking the dialogue-narrative dimension into consideration when interpreting results from a corpus of fiction, it far from invalidates the results from previous cross-linguistic studies. However, one important application of splitting fiction into dialogue and narrative would be to enable studies into the nature of authentic spoken dialogue vs. fictional dialogue, monolingually as well as cross-linguistically.

References

Aijmer, K. (2004). The interface between perception, evidentiality and discourse particle use – using a translation corpus to study the polysemy of SEE. *TradTerm* 10, 246-277.

Alm-Arvius, C. (1993). *The English Verb* See*: A Study in Multiple Meaning.* Göteborg: Acta Universitatis Gothoburgensis.

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

Ebeling, J. (2000). *Presentative Constructions in English and Norwegian. A Corpus-based Contrastive Study*. Oslo: Acta Humaniora.

Egbert, J. & Mahlberg, M. (2017). Fiction — One Register or Two? Narrative and Fictional Speech in Dickens's Novels. *Corpus Linguistics 2017 Conference*, University of Birmingham, 25-28 July 2017 https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper323.pdf.

Johansson, S. (2007). *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam & Philadelphia: Benjamins.

Lefer, M-A. & Vogeleer, S. (eds). (2014). *Genre- and register-related discourse features in contrast*. Amsterdam & Philadelphia: Benjamins.

Neumann, S. (2014). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: De Gruyter Mouton.

Øhman, B.I. (2006). *An SFG perspective on the polysemy of* see*: A corpus-based contrastive study*. Unpublished MA thesis, University of Oslo.

Stubbs, M. & Barth, I. (2003). Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language* 10(1), 61-104.

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

The use of contrastive markers in English policy speeches: A corpus-based cross-modality comparison of *but* and *however* in interpreted and non-interpreted language

Jun Pan
Hong Kong Baptist University
janicepan@hkbu.edu.hk

Policy speeches constitute one of the most important means for the general public to access a government's official policies (Schäffner & Bassnett 2010). The delivery and interpreting of such speeches thus require a high level of pragmatic competence in order to convey the message and underlying attitude in an accurate way (Pan & Wong 2015a, 2015b, forthcoming). In this regard, the appropriate rendition of pragmatic markers (PMs), "the linguistically encoded clues which signal the speaker's potential communicative intentions" (Fraser 1996: 168), becomes significant. Through a corpus-based survey of the use of different PMs in interpreted and non-interpreted political speeches, Pan & Wong (2015a, 2015b) found that contrastive markers (CMs), a subset of PMs, were treated differently than the other types of PMs in interpreted language, including syntactic markers (e.g. "*I know*", "*I think*"), lexical markers (e.g. "*actually*", "*kind of*", "*sort of*", "*then*"), and elaborative markers (e.g. "*above all*"). Whereas the aforementioned PMs were found to be underused in interpreted language, the use of CMs seems to be more complicated: the CM "*however*" was found to be overused while "*but*" and "*instead of*" were underused. In addition, CMs form a special type of PMs in that they signal "the utterance following is either a denial or a contrast of some proposition associated with the preceding discourse" (Fraser: 1996: 187). Despite the pragmatic significance of CMs in policy speeches, little research has been done to show how CMs are used and should be rendered in policy speeches.

This study aims to investigate the use of CMs in interpreted and non-interpreted policy speeches across different modalities (monologues and dialogues) in English, a lingual franca used in international politics (Breiteneder 2009). In this study, interpreted language refers to the language produced via the mediation of interpreters, whilst non-interpreted language refers to the language used in the original speeches, without the mediation of interpreters. Chinese-English interpreting was investigated in the study since Chinese is regarded as implicit and significantly different from English at the pragmatic level (Gu 1992). Two CMs, "*but*" and "*however*", were chosen for analysis since there were the most frequently used CMs in policy speeches (Pan & Wong 2015a, 2015b, forthcoming). In particular, "*but*" signals denial of expectation and contrast (Blakesmore 1989), whereas "*however*" tends to be more subtle in degree and suggest concession (Quirk et al. 1985), the closing of a topic (of a digression) or reintroduction of a prior topic (Bublitz 1988).

The study centers on the following four research questions:
1) What are the differences (if any) in the use of the two CMs in policy speeches interpreted from Chinese to English, and those delivered in English?
2) What are the differences (if any) in the contextual high frequency content words before and after the two CMs in policy speeches interpreted from Chinese to English, and those delivered in English?
3) What are the differences (if any) in the use of the two CMs in policy speeches delivered in a monologue mode as compared to those delivered in a dialogue mode?
4) What are the differences (if any) in the contextual high frequency content words before and after the two CMs in policy speeches delivered in a monologue mode as compared to those delivered in a dialogue mode?

Four corpora were built for the purpose of the study: a parallel corpus consisting of PRC's Premiers' Reports on the Work of the Government and their interpreted texts in English (the PRC corpus), a parallel corpus including Hong Kong SAR's Policy Addresses delivered by its Chief Executives and their interpreted texts in English (the HK SAR corpus), and two comparable corpora made up of the State Opening of Parliament speeches delivered by the Queen of the United Kingdom (the UK corpus) and the State of the Union Addresses delivered by the Presidents of the United States (the US corpus). Speeches delivered at the follow-up press conferences / parliamentary debates of these policy speeches were also collected as they represent similar text types delivered in a dialogue instead of monologue mode. The raw data of the corpora were mostly collected from government official websites, which often provide videos and sometimes edited transcriptions of these speeches. Manual checking (when there was an existing transcription) and transcription (when no transcription was provided) were done to make sure that the texts included in the corpora were complete and consistent.

The Chinese parts of the parallel corpora were processed with word segmentation, using the software *SegmentAnt* (Anthony 2015). They were aligned at the paragraph level using *AntPConc* (Anthony 2014). In addition, *AntConc* (Anthony 2017) was employed for the corpus analysis. All corpora were CM annotated.

Some preliminary analyses were run with the HK SAR corpus, the UK corpus and the US corpus at this stage. Findings of these preliminary analyses show that whilst "*however*" was used with a higher frequency (calculated by average frequency per 10,000 word tokens) in the interpreted policy speeches than the non-interpreted ones in general, "*but*" was used with a much lower frequency in the non-interpreted speeches (Research Question 1). In addition, the collocates of "*but*" seem to be consistent in both the interpreted and non-interpreted policy speeches (Research Question 2). The high frequency content words before "*but*" often pointed to the region/country, whilst those after the CM tend to be its people. The interpreted language, nevertheless, featured more varieties of collocates, as compared to the non-interpreted language. When it comes to the use of "*however*", both the interpreted and non-interpreted policy speeches featured more civil instead of government related topics, and it was especially so in non-interpreted policy speeches (Research Question 2). Further comparisons with its source texts in Chinese will be done to show whether the differences are due to the influence of Chinese or the interpreting process itself, which is cognitively constraining and might impact the use of such markers.

Findings of the study will provide insights into how CMs perform in interpreted and non-interpreted language across different modalities. They will shed light on the training of pragmatic strategies in both policy speech delivery and interpreting.

References

Anthony, L. (2014). *AntPConc* (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Anthony, L. (2015). *SegmentAnt* (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Anthony, L. (2017). AntConc (Version 3.5.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Blakesmore, D. (1989). Denial and contrast: A relevance theoretic account of "but". *Linguistics and Philosophy 12*(1), 15-37.

Breiteneder, A. (2009). English as a lingua franca in Europe: An empirical perspective. *World Englishes 28*(2), 256-269.

Bublitz, W. (1988). *Supportive fellow-speakers and cooperative conversations*. Amsterdam: John Benjamins.

Fraser, B. (1996). Pragmatic markers. *Pragmatics 6*(2), 167-190.

Gu, Y. G. (1992). Pragmatic politeness and culture. *Foreign Language Teaching and Research 4*, 30-32.

Pan, J. & Wong, B. T. M. (2015a). Investigating pragmatic markers in interpreted political speeches from Chinese to English: A preliminary study. In *TSCI 2015 International Conference, "Found in Translation" – Translations are the Children of Their Time,* Bucharest, Romania, 10 September, 6.

Pan, J. & Wong, B. T. M. (2015b). Pragmatic markers in interpreted political discourse: A corpus-driven study. In *International Conference on Corpus Linguistics and Technology Advancement 2015* (CoLTA 2015), Hong Kong, China, 16-18 December, 48.

Pan, J. & Wong, B. T. M. (forthcoming). A corpus-driven study of contrastive markers in Cantonese–English political interpreting. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Schäffner, C. & Bassnett, S. (2010). Politics, media and translation: Exploring synergies. In C. Schäffner & S. Bassnett (eds). *Political Discourse, Media and Translation*. Newcastle: Cambridge Scholars Publishing, 1-31.

Language contact through translation: The effect of explicitness in English-Chinese translation on original Chinese texts

Shuangzi Pang[1], Kefei Wang[2]
Shanghai Jiao Tong University[1], Beijing Foreign Studies University[2]
Melody2459@hotmail.com, kfwang@bfsu.edu.cn

Translation has played a crucial role in the history of language, knowledge and culture as a subject, which on the one hand, is interfered by source languages and on the other hand, may also impact on target texts. But the influence of translation on target language has long been neglected due to the limitation of measuring means. During the end of 1990s, the marriage of corpus linguistics and descriptive translation studies gave rise to a considerable body of empirical research between translated texts and original texts, which makes it possible to examine translation as a site of language contact through written texts. However, so far, translation as the site of language contact has not been investigated in depth either in linguistic area or via translation studies. Moreover, research in this area has, until recently been confined to the comparison between related European languages, and it is of vital importance to find evidence from genetically distinct language pairs such as English and Chinese.

In the English-Chinese translation field, previous work on language contact through translation has generally been conducted from two perspectives: vocabulary and syntax which are rather fragmentary or dealt with a passing. In order to test the phenomenon more systematically, this paper purports to examine whether the "explicitness" in English-Chinese literary translations might have an effect on target texts in order to observe the diachronic relationship between translated texts and non-translated texts in three time frames: 1930s, 1960s and 1990s. Meanwhile, source language interference and implicitation will both be taken into account in order to compensate for the weakness of the previous research. The study will address the following questions:

1) Does the explicitness in English-Chinese translated texts and non-translated texts show a diachronic change over the three sampling periods?
2) Do the translated texts and non-translated texts correlate with each other in the three time frames? How does the explicitness in translated texts impact on the target texts diachronically?
3) Are the adversative conjunctions in translated texts interfered by the source language? How do explicitation and implicitation change in E-C translation over the three periods?

These questions will be answered by compiling English-Chinese Diachronic Composite Corpora, which incorporate a diachronic English-Chinese parallel corpus (TC), a comparable diachronic Chinese corpus (CC), and a "pure" Chinese reference corpus (RC). The TC and CC have been divided into three corpora separately, containing the texts from 1930s, 1960s, 1990s which roughly correspond to the sample periods of FLOB. The corpora (approximately 10 million words) have been aligned sentence to sentence in parallel section, and annotated with The University of Pennsylvania Treebank Tag-set for English texts and ICTCLAS in Chinese Academy of Sciences for Chinese texts separately. Part of the corpora has been annotated semi-automatically with the software of BFSU (Beijing Foreign Studies University) Qualitative Coder in order to examine the concordances of explicitating and implicitating shifts in three periods.

Initially, the terms of "explicitation" and "implicitation" are clarified and the conjunctions are categorized into three types among which adversative conjunctions, which are the main topic of the presentation. The diachronic corpora have been analyzed using a three-step method. The change of explicitness in translated texts over the three sampling periods is examined firstly under the scrutiny of parallel corpora in order to see if the frequency development is due to source language interference. The comparable corpora are analyzed to determine whether the change in translated Chinese texts also happen in the original Chinese texts of the same period or the next period. Finally, the reference corpora are employed to triangulate the data with a view to determining the extent of the change of explicitness in original Chinese language. Correspondence analysis is employed to find the clusters of conjunctions so as to unveil the preference of translational shifts in the three periods and measure the distance between translated and original Chinese texts diachronically.

It is found that:

1) On the whole, translated Chinese texts have changed concomitantly with original Chinese texts in the frequencies of adversative conjunctions, i.e. fewer adversative words have been employed over time; in some particular period, however, the adversative conjunctions employed in the two types of texts became divergent (see Table 1).

2) The adversative conjunctions in the translated texts (TC3) and original Chinese texts of the third period (CC3) cluster closer to each other than TC1&CC1 and TC2 & CC2, according to correspondence analysis (see Figure 1), which indicates that translated Chinese texts and original Chinese texts in the three periods are interrelated, but the correlation changed perceptibly over the three sampling points.

3) The occurrences of the equivalents to the original English in translated Chinese texts have increased, showing that source language interference in translated Chinese texts has ascended over the three sampling periods; the decrease of explicitation and implicitation in translations suggests that literal translation has been preferred to free translation (see Table 2). Additionally, the ratio of explicitation and implicitation in translation changed across time.

	1927-1937	1956-1962	1987-1997	R1	R2
TC	6. 27	6.57	5.94	1.05	0.95
CC	6.16	3.39	2.89	0.55	0.85
RE	1.02	1.94	1.98	——	——

Table 1. Diachronic change of adversative conjunctions in TC and CC

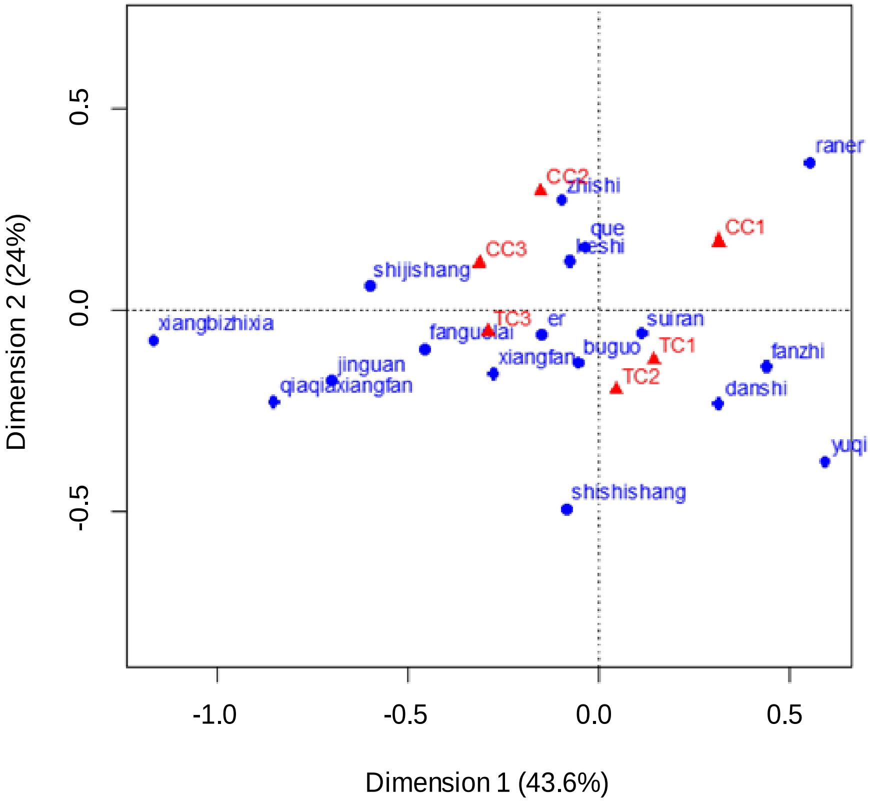


Figure 1. Distance between translated Chinese texts (TC) and original Chinese texts (CC) in the use of adversative conjunctions over the three periods

	1927-1937	1956-1962	1987-1997	R1	R2
explcitation	0.19	0.18	0.14	0.95	0.77
implicitation	0.35	0.18	0.16	0.51	0.89
correspondence	0.6	0.63	0.7	1.05	1.1
Explici/implici ratio	0.54	1	0.875	1.85	0.875

Table 2. Diachronic shifts of explicitation and implicitation in parallel corpora

Based on this, the study explores whether "translation universals" still exist through diachronic study in Chinese, and delineates to what extent translation has influenced modern vernacular Chinese development. The paper further aims to explore how the diachronic composite corpora can be fruitfully used to study the relation between translation and language change, in terms of the methodology concerned and especially the relation between English-Chinese translation and Chinese language development.

References

Amouzadeh, M. & House, J. (2010). Translation as a Language Contact Phenomenon: The Case of English and Persian Passives. *Languages in Contrast* 10(1), 54-75.

Baker, M. (1996). Corpus-based Translation Studies: The Challenges that Lie Ahead. In H. Somers (ed.) *Terminology, LSP and Translation*. Amsterdam: John Benjamins, 175-86.

Baumgarten, N. (2007). Converging conventions? Macrosyntactic Conjunction with English and German. *Text Talk* 27, 139-170.

Bennett, K. (2010). Academic Discourse in Portugal: A Whole Different Ballgame? *Journal of English for Academic Purposes* 9 (1), 21-32.

Bisiada, M. (2013). *From Hypotaxis to Parataxis: An investigation of English-German Syntactic Convergence in Translation*. PhD thesis. Manchester University.

Blake, N. F. (1992). Translation and the History of English. In M. Rissanen, O. Ihalainen, T. Nevalainen & I. Taavitsainen (eds). *History of Englishes: Methods and Interpretations in Historical Linguistics*. Berlin: De Gruyter, 3-24.

Blum-Kulka, S. (1986). Shifts of Cohesion and Coherence in Translation. In J. House & S. Blum-Kulka (eds). *Interlingual and Intercultural Communication. Tübingen:* Gunter Narr, 290-305.

Braunmuller, K. & House, J. (2009). *Convergence and Divergence in Language Contact Situations*. Amsterdam: John Benjamins.

Dayrell, C. (2005). *Investigating Lexical Patterning in Translated Brazilian Portuguese: A Corpus-Based Study*. PhD Thesis. University of Manchester.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

House, J. (2011). Using Translation and Parallel Text Corpora to Investigate the Influence of Global English on Textual Norms in Other Languages. In A. Kruger, K. Wallmach & J. Munday (eds). *Corpus-Based Translation Studies: Research and Applications*. London: Continuum International Publishing Group, 187-208.

Hu, K. (2005). On the Impact of the Historical Text of English-Chinese Dictionary on the Modernization of the Chinese Language. *Foreign language and Foreign language teaching* 37(3), 57-60.

Kranich, S., Becher, V. & Hoder, S. (2011). A Tentative Typology of Translation-induced Language Change. In S. Kranich, V. Becher, S. Hoder & J. House (eds). *Multilingual Discourse Production.* Amsterdam: John Benjamins Publishing Company, 11-43.

Laviosa, S. (2002). *Corpus-based Translation Studies: Theory, Findings and Applications*. Amsterdam: Rodopi.

Malamatidou, S. (2013). *Translation and Language Change with Reference to Popular Science Articles: The Interplay of Diachronic and Synchronic Corpus-Based Studies*. PhD thesis. Manchester University.

McEnery, T. & Xiao, R. (2007). Parallel and comparable Corpora: What is happening? In G. Anderman & M. Rogers (eds). *Incorporating Corpora: The Linguist and the Translator.* Bristol: Multilingual Matters, 18-31.

Xiao, R. & Wei Naixing, W. (2014). Translation and contrastive linguistic studies at the interface of English and Chinese: Significance and implications. *Corpus Linguistics and Linguistic Theory* 10(1), 1-10.

Vinay, J.-P. & Darbelnet, J. (1958/1995). *Comparative Stylistics of French and English: A Methodology for Translation.* Amsterdam & Philadelphia: John Benjamins Publishing Company.

Wang, K. (2002). Influence of Modern Translation on Chinese. *Foreign language teaching and research* 34(6), 458-463.

Weinreich, U. (1953/1974). *Language in Contact: Findings and Problems*. The Hague: Mouton.

Analyzing participant inclusion in Spanish, French and English passive-like structures: The influence of pragmatic-discursive factors

Emeline Pierre
Université catholique de Louvain
emeline.pierre@uclouvain.be

This paper aims to analyze the use of different structures that allow for expressing the passive voice in Spanish, French and English in formal and informal registers. Besides taking taking into account intrinsic characteristics of the structures, the study examines in detail pragmatic features and register variation. While comparative work on (some of) these passive structures has been carried out from a typological perspective (Siewierska 1984; Rousseau 2000; Siewierska & Papastathi 2011; Gast & van der Auwera 2013) or contrasting specific constructions in two languages (e.g. Espinoza 1997; Haverkate 2004), a contrastive study of the panorama of passives in different languages and registers is lacking. More precisely, this paper focuses on passive constructions based on a reflexive pronoun (*se* in Spanish and French), man-impersonals (*on* in French) and numeral-based indefinite pronouns (*uno* in Spanish and *one* in English). To conduct my study, I use both written and oral data extracted from comparable informal corpora (Yahoo Questions & Answers for written data and ESLORA2 for Spanish oral data, CORALROM for French oral data and BNC for English oral data) and comparable formal corpora (Wikipedia for written data and Europarl for oral data). Analyzing corpus-based data, I aim to evaluate the potential impact of the inclusion/exclusion of participants on the selection of passive-like structures and to determine to what extent formally similar structures function differently across the languages under study.

The languages selected for the present study pertain to the Indo-European language group and may share some grammatical properties. However, it is hypothesized that these languages do not always resort to the same structures and apparently similar constructions may convey slightly different meanings, hence the importance and interest of establishing and examining a selection of passive related structures. Our two research questions will thus be: (1) how may the linguistic environment affect the inclusion/exclusion of participants and what is the impact on the selection of passive-like structures in Spanish, French and English? and (2) is there a possible influence of the genre on the inclusion/exclusion of participants in the expression of passive voice?

Taking into account the linguistic environment, it is expected that linguistic factors such as person references (1), adverbial constructions (2) or prepositional phrases (3) in the nearby context have a crucial importance in determining which participants are included in the interpretation of the passive-like structure.

(1) *(…) do you want us to confirm your desire is possible or do you want the real answer from real aviators? {S} Part of maturity is being able to recognize that* ***one's*** *desires may not be obtainable*.

(2) *Generalmente* ***se acepta*** *que el fruto se desarrolla posteriormente a la fertilización, sin embargo, esto no es necesariamente cierto siempre, (…).*
‘Usually **it is accepted** that fruits grow after fertilization, however, this is not always necessarily sure.

(3) *Comment* ***fait-on*** *dans les pays nordiques ou au Canada ou les températures sont bien plus basses et la neige plus abondante ?*
‘How **do they** do in Nordic countries or in Canada where temperatures are far lower and snow heavier.’

Preliminary results show that when there is a person reference in the nearby context the constructions under study display various similarities and differences both across languages and across registers. The Spanish *se* is more productive than its French counterpart is and functions slightly differently. The pilot analysis reveals that, in the Corpus Yahoo Questions & Answers, the Spanish *se* occurs more frequently with a person reference in its nearby context than the French construction with *se*. In addition, both constructions display variations in participant inclusion. This may suggest that these formally similar structures occupy a different position on the continuum of passive-like structures.

My first findings also suggest that the feature 'person reference in the context' shows a similar behavior in informal written data and a different one in formal written data. For example, in Spanish informal data, constructions with *se* and *uno* both tend to include a person reference in their context and to have the possibility to include a different range of participants. In formal data, the Spanish *se* tends to refer to a person outside the communicational situation while *uno* is often interpreted as maximally inclusive. The pilot analysis reveals that 'person reference in the context' is not the only factor accounting for participant inclusion. The presence of an adverbial construction may also have an impact when determining which participant is included in the passive-like structure. When the latter has a generic interpretation, the *se* construction (both in French and in Spanish), accompanied by a generalizing adverb, tends to be favored. The use of an adverb would be more specific of the Spanish language.

These first analyses furthermore suggest that the phenomenon of participant inclusion in passive voice constructions will be more typical of informal registers. I expect the oral data to have a similar behavior, the presence of a person reference in the context of the passive-like structures being strengthened. Ultimately, this investigation suggests that the studied structures have different positions on the panorama of passive voice constructions, some receiving a more impersonal interpretation than others. Such a large panorama of passive-like structures may help linguistic experts evolving in a global environment to have a better understanding of the subtleties and differences between the languages under study.

References

Espinoza, A. M. (1997). Contrastive analysis of the Spanish and English passive voice in scientific prose. *English for specific purposes* 16(3), 229-243.

Gast, V. & van der Auwera, J. (2013). Towards a distributional typology of human impersonal pronouns, based on data from European languages. In D. Bakker & M. Haspelmath (eds). *Languages across boundaries. Studies in memory of Anna Siewierska*. Berlin: De Gruyter Mouton, 119-158.

Haverkate, Henk. (2004). Gramática y pragmática: categorías desfocalizadores en español. *Spanish in Context* 1(1), 21-40.

Rousseau, André. (2000). Formation et statut du passif. Comparaison typologique entre langues romanes et langues germaniques. In L. SchØsler (ed.) *Le Passif. Actes du colloque international Institut d'Etudes Romanes Université de Copenhague du 5 au 7 mars 1998*. Études Romanes 45. Copenhagen: Museum Tusculanum Press, 117-133.

Siewierska, Anna. (1984). *The passive. A comparative linguistic analysis*. London: Croom Helm.

Siewierska, A. & Papastathi, M. (2011). Towards a typology of third personal plural impersonals. *Linguistics* 49(3), 575-610.

A corpus-based contrastive analysis of the Dutch adjectival *-s* ending. Deflexion or refunctionalization?

Dirk Pijpops[1,2], Freek Van de Velde[2]
Research Foundation Flanders (FWO)[1], University of Leuven[2]
dirk.pijpops@kuleuven.be, freek.vandevelde@kuleuven.be

In this study, we contrast the Moroccan-Dutch ethnolect with the language use of full native speakers within the framework of Contrastive Interlanguage Analysis (Granger 2015). Our focus will be on their realization of Dutch adjectival morphology. Language users of the Moroccan-Dutch ethnolect may be creatively restructuring Dutch morphology in a number of novel ways. In particular, the adjectival *-e* inflection has drawn scholarly attention (Van de Velde & Weerman 2014). Here, it is argued that these language users are revitalizing a seemingly defunct inflection system by discarding a number of synchronically unmotivated exceptions. The *-e* ending may then acquire new functions as (i) a marker of attributive modification and (ii) a boundary marker between the modification and determination zones in the noun phrase. The *-e* ending is not the only remnant of the once elaborate Dutch adjectival inflection system, however. The so-called partitive genitive construction also harbors an adjectival *-s* ending, that, like the *-e*, alternates with a zero ending, as in (1) versus (2) (Haeseryn et al. 1997: 421, for the contexts in which either form is used in Present-day Dutch, see Pijpops & Van de Velde 2014).

(1)	*de*	*hijab*	*is*	***iets***	***moois***	*wat*	*door*	*Marokkaanse*	*wijven*	*helemaal*	*verpest*	*is*
	the	hijab	is	**something**	**beautiful-GEN**	that	by	Moroccan	women	totally	ruined	is

‘The hijab is something beautiful that is totally ruined by Moroccan women.’

(2)	*Is*	*dat*	***iets***	***verkeerd***
	is	that	**something**	**wrong-Ø**

‘Is that something wrong?’

This *-s* ending is one of the few surviving remnants of the Dutch genitive case, more specifically the partitive genitive: hence the name of the construction. The partitive genitive construction is a combination of an indefinite pronoun or numeral with a postmodifying adjectival phrase, although the exact theoretical architecture of the construction is still very much up for debate (Schultink 1962: 62; Kester 1996; van Marle 1996; Broekhuis & Strang 1996; Hoeksema 1998; Booij 2010: 223-228; Broekhuis 2013: 419-461). We then ask whether and, if so, how the language users of the Moroccan-Dutch ethnolect differs from full native language use in the utilization of this adjectival *-s* ending. This may be a difference in absolute numbers, but may also pertain to the number and choice of the factors that determine the appearance of the *-s* ending. There are four possible options:

(i) Like the *-e* ending, the users of Moroccan Dutch generalize the *-s* ending to all instances of the partitive genitive, thereby refunctionalizating this remnant of the Dutch case system as a transparent and reliable construction marker (cf. Booij 2010: 223-228).

(ii) The users of Moroccan Dutch generalize the zero ending, thereby ridding their language of an obsolete fossil from bygone times. This would be a continuation of the deflexion trend apparent in the development of Dutch (van der Horst 2013). The resulting state would be akin to English, where only the zero ending is used.

(iii) The users of Moroccan Dutch employ both the *-s* and zero ending in exactly the same way as other language users of Dutch, implementing the same factors to determine the choice between both variants. This would indicate that these factors are of a qualitatively different nature than the factors determining the use of the *-e* ending, as the users of Moroccan Dutch apparently do not or cannot dispose of them.

(iv) The users of Moroccan Dutch employ both the *-s* and zero ending in the partitive genitive construction, but in a different way than other language users of Dutch. This would indicate that these language users are creatively adapting their language to cater to new or other needs

To investigate this, we will apply regression modelling to corpus data, as proposed by Gries & Deshors (2014). Gries & Deshors advocate this methodology as a way to fully exploit the potential of Granger's (1996) model of Contrastive Interlanguage Analysis (CIA). As a source of data, we turn to the Moroccorp corpus, which contains chat conversations in the Moroccan-Dutch etnolect and has already proven its value in studies on Dutch adjectival inflection (Ruette & Van de Velde 2013; Van de Velde & Weerman 2014). We extracted a number of possible partitive genitives which we manually filtered, and finally retained 1613 genuine partitive genitive instances. These partitive genitives are contrasted with 765 observations of partitive genitives taken from Netherlandic chat conversations in the ConDiv corpus (Grondelaers et al. 2000), adopted from an earlier study by Pijpops & Van de Velde (2014). The Moroccorp corpus was specifically designed to be commensurable to this subsection of ConDiv (Ruette & Van de Velde 2013: 467-470). Finally, we employed mixed logistic regression modelling to investigate how the realization of partitive genitives differs in both corpora.

This study will shed light on how early L2/2L1 speakers deal with seemingly defunct morphology in Dutch, and hopes to answer the call of Gries and Deshors (2014) for more elaborate statistical methods in CIA research.

References

Booij, G. (2010). *Construction morphology*. Oxford: Oxford University Press.

Broekhuis, H. (2013). *Syntax of Dutch. Adjectives and Adjective Phrases*. Amsterdam: Amsterdam University Press.

Broekhuis, H. & Strang, A. (1996). De partitieve genitiefconstructie [The partitive genitive construction]. *Nederlandse taalkunde* 1(3), 221-238.

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

Gries, S. Th. & Deshors, S. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora* 9(1), 109-136.

Grondelaers, S., Deygers, K., Van Aken, H., Van den Heede, V. & Speelman, D. (2000). Het CONDIV-corpus geschreven Nederlands [The CONDIV-corpus of written Dutch]. *Nederlandse Taalkunde* 5(4), 356-363.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. & van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst [General Dutch Grammar]*. Groningen: Nijhoff.

Hoeksema, J. (1998). Adjectivale inflectie op -s: geen geval van transpositie [Adjectival inflection on -s: not a case of transposition]. In E. Hoekstra & C. Smits (eds). *Morfologiedagen 1996 [Morphology Days 1996]*. Amsterdam: P. J. Meertens-Instituut, 46-72.

Kester, E.-P. (1996). *The nature of adjectival inflection*. Utrecht: LEd.

Pijpops, D. & Van de Velde, F. (2014). A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory*. Published online, ahead of print.

Ruette, T. & Van de Velde, F. (2013). Moroccorp: tien miljoen woorden uit twee Marokkaans-Nederlandse chatkanalen. *Lexikos* 23, 475-556.

Schultink, H. (1962). *De morfologische valentie van het ongelede adjectief in modern Nederlands*. Den Haag: Van Goor.

van Marle, J. (1996). The unity of Morphology: on the interwovenness of the derivational and inflectional dimension of the word. *Yearbook of Morphology 1995*. Dordrecht: Kluwer, 67-82.

van der Horst, J. (2008). *Geschiedenis van de Nederlandse syntaxis [History of Dutch syntax]*. Leuven: Universitaire Pers Leuven.

van der Horst, J. (2013). *Taal op drift. Lange-termijnontwikkelingen in taal en samenleving [Language adrift. Long term developments in language and society]*. Amsterdam: Meulenhoff.

Van de Velde, F. & Weerman, F. (2014). The resilient nature of adjectival inflection in Dutch. In P. Sleeman, F. Van de Velde & H. Perridon (eds). *Adjectives in Germanic and Romance* (Linguistik Aktuell/Linguistics Today). Amsterdam: John Benjamins, 113-145.

Lexis or parsing? A corpus-based study of syntactic complexity and its effect on disfluencies in interpreting

Koen Plevoets[1,2], Bart Defrancq[1]
Ghent University[1], KU Leuven[2]
koen.plevoets@ugent.be, bart.defrancq@ugent.be

Cognitive load is probably one of the most cited topics in research on simultaneous interpreting, but it is still poorly understood due to the lack of proper empirical tests. It is a central concept in Gile's (2009) Efforts Model as well as Seeber's (2011) Cognitive Load Model. Both models invariably conceptualize interpreting as a dynamic equilibrium between the cognitive resources/capacities and cognitive demands that are involved in listening and comprehension, production and memory storage. In cases when the momentary demands exceed the interpreter's available capacities, there is an information overload which typically results in a disfluent or erroneous interpretation. While Gile (2008) denies that his Efforts Model is a theory that can be tested, Seeber & Kerzel (2012) put Seeber's Cognitive Load Model to the test using pupillometry in an experimental interpretation task.

In a series of recent corpus-based studies Plevoets & Defrancq (2016, 2018) and Defrancq & Plevoets (2018) used filled pauses to investigate cognitive load in simultaneous interpreters, based on the widely shared assumption in the psycholinguistic literature that silent and filled pauses are 'windows' on cognitive load in monolingual speech (Arnold et al. 2000; Bortfeld et al. 2001; Clark & Fox Tree 2002; Levelt 1983; Watanabe et al. 2008). The studies found empirical support for increased cognitive load in simultaneous interpreting in the form of higher frequencies of filled pauses. However, the studies also showed that filled pauses in interpreting are caused mainly by problems with lexical retrieval. Plevoets & Defrancq (2016) observed that interpreters produce more instances of the filled pause *uh(m)* when the lexical density of their own output is higher. Plevoets & Defrancq (2018) demonstrated that the frequency of *uh(m)* in interpreting increases when the lexical density of the source text is also higher but it decreases when there are more formulaic sequences. This effect of formulaicity was found in both the source texts and the target texts. Other known obstacles in interpreting, such as the presence of numbers and rate of delivery do not significantly affect the frequency of filled pauses (although source speech delivery rate reached significance in one of the analyses). These results point to the problematic retrieval or access of lexical items as the primary source of cognitive load for interpreters. Finally, in a study of filled pauses occurring between the members of morphological compounds, Defrancq & Plevoets (2018) showed that interpreters produced more *uh(m)*'s than non-interpreters when the average frequency of the compounds was high as well as when the average frequency of the component members was high. This also demonstrates that lexical retrieval, which is assumed to be easier for more frequent items, is hampered in interpreting.

This study critically examines the results of the previous studies by analyzing the effect of another non-lexical parameter on the production of filled pauses in interpreting, viz. syntactic complexity. Subordinating constructions are a well-known predictor of processing cost (cognitive load) in both L1 research (Gordon et al. 1986; Gordon & Luper 1989) and L2 research (Norris & Ortega 2009; Osborne 2011). In interpreting, however, Dillinger (1994) and Setton (1999: 270) did not find strong effects of the syntactic embedding of the source texts on the interpreters' performance. As a consequence, this paper will take a closer look on syntactic complexity and it will do so by incorporating the number of hypotactic clauses into the analysis.

The study is corpus-based and makes use of both a corpus of interpreted language and a corpus of non-mediated speech. The corpus of interpreted language is the EPICG corpus, which was compiled at Ghent University between 2010 and 2013. It consists of French, Spanish and Dutch interpreted speeches in the European Parliament from 2006 until 2008, which are transcribed according to the VALIBEL guidelines (Bachy et al. 2007). For the purposes of this study a sub-corpus of French source speeches and their Dutch interpretations is used, amounting to a total of 140,000 words. This sub-corpus is annotated for lemmas, parts-of-speech and chunks (Van de Kauter et al. 2013), and it is sentence-aligned with WinAlign (SDL Trados WinAlign 2014). The corpus of non-mediated speech is the sub-corpus of political debates of the Spoken Dutch Corpus (Oostdijk 2000). The corpus was compiled between 1998 and 2003, and it is annotated for lemmas and parts-of-speech. The political sub-corpus contains 220,000 words of Netherlandic Dutch and 140,000 words of Belgian Dutch.

The data are analysed with a Generalized Additive Mixed-effects Model (Wood 2017) in which the frequency of the disfluency *uh(m)* is predicted in relation to delivery rate, lexical density, percentage of numbers, formulaicity and syntactic complexity. Delivery rate is measured as the number of words per minute, lexical density as the number of content words per utterance length, percentage of numbers as the numbers of numbers per utterance length and formulaicity as the number of n-grams per utterance length. The new predictor, syntactic complexity, is measured as the number of subordinate clauses per utterance length. Because all five predictors are numeric variables, their effects are modelled with smoothing splines which automatically detect potential nonlinear patterns in the data. The observations are at utterance-level and are nested within the speeches, so the possible between-speech variation is accounted for with a random factor.

The preliminary results confirm the hypothesis: while lexical density and formulaicity show similar effects to what is reported in previous research (i.e. lexical density is positively associated with the frequency of *uh(m)* and formulaicity is negatively associated with it), the syntactic complexity of the source text is 'border-significant' and the syntactic complexity of the target is non-significant. There are some sporadic differences among certain types of subordinate clauses, but the general conclusion is indeed that syntactic complexity is not such a strong trigger of cognitive load in interpreting in comparison to lexically-related factors. That calls for a model of interpreting in which depth of processing plays only a marginal role.

References

Arnold, J. E., Wasow, T., Losongco, A. & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76, 28-55.

Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.-C. & Wilmet, R. (2007). *Conventions de transcription régissant les corpus de la banque de données VALIBEL*. www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.PDF.

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.

Clark, H. H. & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition* 84, 73-111.

Defrancq, B. & Plevoets, K. (2018). Over-*uh*-load. Filled pauses in compounds as a signal of cognitive load. In M. Russo, C. Bendazolli & B. Defrancq (eds). *Making way in corpus-based interpreting studies*. Berlin: Springer, 43-63.

Dillinger, M. (1994). Comprehension during interpreting: What do interpreters know that bilinguals don't? In S. Lambert & B. Moser-Mercer (eds). *Bridging the gap: Empirical research in simultaneous interpretation*. Amsterdam: John Benjamins, 155-189.

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.

Gordon, P. A. & Luper, H. L. (1989). Speech disfluencies in nonstutterers. Syntactic complexity and production task effects. *Journal of Fluency Disorders* 14, 429-445.

Gordon, P. A., Luper, H. L. & Peterson, H. A. (1986). The effects of syntactic complexity on the occurrence of dislfuencies in 5 year old nonstutterers. *Journal of Fluency Disorders* 11, 151-164.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition* 14, 41-104.

Norris, J. M. & Ortega, L. (2009). Toward an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics* 30 (4), 555-578.

Oostdijk, N. (2000). The Spoken Dutch Corpus: Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhaouer (eds). *Proceedings of the Second International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), 887-894.

Osborne, J. (2011). Fluency, complexity and informativeness in native and non-native speech. *International Journal of Corpus Linguistics* 16 (2), 276-298.

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

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

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L. & Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal* 3, 103-120.

Watanabe, M., Hirose, K., Den, Y. & Minematsu, N. (2008). Filled pauses as cues to the complexity of up-coming phrases for native and non-native listeners. *Speech Communication* 50, 81-94.

SDL Trados WinAlign (2014). *SDL Trados WinAlign Tutorial*. www.translationzone.com/resources/downloads/winalign-tutorial.html.

Wood, S. N. (2017). *Generalized additive models: An introduction with R. Second edition*. Boca Raton: Chapman & Hall/CRC.

Obituaries in English-Portuguese translation: A corpus-based study

Rozane Rebechi, Márcia Moura da Silva
Federal University of Rio Grande do Sul
rozanereb@gmail.com, marciamouras@hotmail.com

The objective of this study is to verify whether – and to what extent – a comparable Brazilian Portuguese and North-American English corpus of obituaries can help with the task of retrieving equivalents for terms and phraseologies which are characteristic of this genre, and discuss the implications of the findings for translator training.

Despite affecting everyone, everywhere, death is faced differently across cultures (Eid 2002), and according to Bates et al. (2009: 2), "The way a culture chooses to commemorate its dead reflects a great deal about the character and nature of that culture". When we compare Brazil and the United States, huge differences emerge. Funeral duration is one example. While North-American funerals may last several days, inasmuch as they involve, among other stages, mourning the deceased, celebrating their life, and offering support to the family, Brazilians tend to proceed with burial or cremation in no more than 24 hours after death. Another difference is the availability and content of obituaries – understood here both as edited biographies of notorious deceased and paid death notices – in both cultures. After analyzing the essence of contemporary North-American obituaries, Bates et al. (2009) concluded that these texts draw audience to newspapers, corroborating Starck's claim: "In the English-speaking world, a newspaper of quality hardly seems complete these days without a regular obituary page" (Starck 2006: x). In Brazil, on the other hand, the theme is far from popular. Here obituaries are mostly dedicated to famous people in the form of news articles written by journalists and published in some newspapers and magazines under sections such as 'deaths', 'daily', 'fun and art' and 'sports', depending on the area in which the deceased gained fame, and paid death notices are usually short and restricted to giving information about the (bygone) funeral.

Such particularities certainly have an impact on the standard terminology used in obituaries. Unsurprisingly, rendering North-American obituaries into Portuguese resulted in a challenging task for undergraduate students in our institution. Among the difficulties, we identified (i) lack of terminology standardization - 'visitation' was rendered as '*cerimônia*', '*velório*' and '*visitação*', while '*missa*', '*cortejo*' and '*serviço*' were suggested equivalents for 'service'; (ii) literal translation (Chesterman 1997) resulting in non-idiomatic structures in the target language – 'is survived by' was rendered as '*é sobrevivido(a) por*', and 'is/was preceded (in death) by' and 'is/was predeceased by', as '*foi precedido (na morte) por*'; (iii) inadequate translation which leads to false information – '*esposa de x anos*' corresponds, in Portuguese, to the widow's age, not to the duration of the marriage, as appeared in the source text ('wife of x years').

Following Philip (2009), who claims that comparable corpora can reveal the terminology and phraseologies used naturally in the source language, in addition to helping in the identification of discrepancies between textual types produced in different languages and cultures, we selected 200 obituaries in each language – 100 edited texts and 100 paid notices, published between 2016 and 2017 in major Brazilian and North-American newspapers, accounting for 54,415 and 90,402 tokens, respectively. Both subcorpora were analyzed semi automatically with WordSmith 6.0 (SCOTT 2012). After comparing our study corpus to reference corpora (American National Corpus for the English subcorpus and LacioRef for the Portuguese subcorpus), we manually analyzed the single and compound keywords retrieved, which revealed both linguistic and cultural differences.

In the English subcorpus, we identified keywords related to the several stages of North-American funerals, such as 'graveside service', 'memorial service', 'celebration of life', 'funeral mass', 'wake' and 'viewing'. Conversely, the analysis of the Brazilian Portuguese subcorpus confirms that death rites are usually limited to '*velório*' – an often public funeral ceremony in which the coffin is put on public display to allow relatives, friends and other interested parties to honor the memory of the deceased, followed by '*enterro/sepultamento*' (burial).

As Franco Aixelá (1996: 53) points out, “cultures create a variability factor the translator will have to take into account”. When rendering North-American obituaries in Brazilian Portuguese, not only do translators need to be aware of the differences between North American and Brazilian ceremonies, which give rise to different terms, but they also need to consider cultural aspects, which may have direct impact on the choice of equivalents. For example, a common phraseology in North-American obituaries is ‘[name of the deceased] is survived by’, which corresponds, in Portuguese, to ‘[name of the deceased] *deixa*’ (third person singular of the verb ‘leave’). This example demonstrates that North-American and Brazilian cultures have a different perspective of death: while the former uses the passive voice, focusing attention on the survivors, the latter uses the active voice, with the focus on the deceased.

Another common practice in the North-American obituaries is the request of donations to various institutions, revealed by the recurring phraseology ‘in lieu of flowers donations can/may be made to’. A possible correspondence is absent from the Portuguese subcorpus. The keywords also showed that in Brazilian Portuguese obituaries mention to survivors is usually restricted to close family members, such as offspring and spouse, whereas in North-American counterparts a whole list of relatives is mentioned, including the ones who have already passed away, introduced by the phraseologies ‘was predeceased by’ and ‘was preceded in death by’, whose possible equivalents in Portuguese are not statistically representative.

Despite seemingly small, the comparable specialized corpus led to interesting results in terms of terminology and phraseology, since the corpus is representative of a specialized area (cf. Koester 2010). Nevertheless, the study corpus has limitations as a source of terminology and phraseology retrieval, since not only are obituaries more common in the North-American culture, but they are also longer. Recurring terms and phraseologies also highlight distinctions in both languages/cultures, some of which are hardly rendered in Brazilian Portuguese, since we do not have all the referents. However, in accordance with Loock & Lefebvre-Scodeller (2014), we believe that cultural differences unveiled by obituaries may be used with translation trainees as a limited source for terminology retrieval, with special emphasis to cultural differences.

References

Bates, A., Monroe, I. & Zhuang, M. (eds). (2009). *The state of the American obituary*. Medill, Nov. 30. Available at: https://www.ianmonroe.com/wp-content/uploads/2009/10/StateOfTheAmericanObituary_Nov2009.pdf. (Accessed Jan. 22, 2018).

Chesterman, A. (1997). *Memes of Translation: The Spread of Ideas in Translation Theory*. Amsterdam & Philadelphia: John Benjamins.

Eid, M. (2002). *The world of obituaries: gender across cultures and over time*. Detroit: Wayne State University Press.

Franco Aixelá, J. (1996). Culture-specific items in translation. In R. Álvarez & M. C. A. Vidal (eds).. *Translation power subversion*. Clevedon: Multilingual Matters, 52-78.

Koester, A. (2010). Building small specialised corpora. In A. O’Keeffe & M. McCarthy (eds). *The Routledge handbook of Corpus Linguistics*. New York: Routlege, 66-79.

Loock, R. & Lefebvre-Scodeller, C. (2014). Writing about the dead: a corpus-based study on how to refer to the deceased in English vs French obituaries and its consequences for translation. In M. Garant (ed.) *Current trends in translation teaching and learning*. Helsinki: University of Helsinki, 115-150.

Philip, G. (2009). Arriving at equivalence: making a case for comparable general reference corpora in translation studies. In A. Beeby, P. R. Inés & P. Sánchez-Gijón (eds). *Corpus use and translating*. Amsterdam & Philadelphia: John Benjamins, 59-73.

Scott, M. (2012). *Wordsmith Tools version 6.0*. Oxford: Oxford University Press.

Starck, N. (2006). *Life after death: the art of the obituary*. Melbourne: Melbourne University Press.

A bilingual parallel corpus (EN-ES) of texts translated by students as a tool for research purposes

Juan Pedro Rica Peromingo, Arsenio Andrades Moreno, Jorge Braga Riera,
Nava Maroto García, Ángela Sáenz Herrero, Sara Martínez Portillo
Universidad Complutense de Madrid
juanpe@ucm.es, arsenio.andrades@uca.es, jbragariera@filol.ucm.es,
mariadelanava.maroto@upm.es, angsaenz@ucm.es, samart01@ucm.es

Nowadays, corpus linguistics has become a key research methodology for Translation Studies (Granger & Lefer 2017) which broadens the scope of cross-linguistic studies. Corpus linguistics has had a huge impact on translation theory and practice, and is a major tool to identify errors and analyze translation strategies in a bilingual corpus. It is a research paradigm with a strong potential to significantly contribute to translation teaching and language learning (Johansson 2007; Rica et al. 2014; Rica & Braga 2015; Rica in press 2018). Although most studies on translation are based on the observation of different aspects of the translation process of professional translators, in this case the approach is changed to focus on learners with little or no experience in order to study, at an early stage, general mistakes and to improve the translational competence of the students.

Led by Sylviane Granger and Marie-Aude Lefer of 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 Europe 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) an indirect (L1>L2) translations, from a great variety of text types, genres and registers in a wide range of languages. Some of the first questions that arise upon considering the utility of this corpus are: Will there be any significant differences among the different translation strategies employed in the translations depending on the genre of the texts? Will the students with a higher level of English resort to a wider range of translation strategies? Will the students translating different text genres resort to different strategies depending on the text type? It seems from preliminary results gathered on the first analysis stages of this process that the answer to all these questions is an affirmative one, which leads us to the proposition of further questions such as: What are the mechanisms more frequently resorted to? Are there any mistakes in the recurrence of the use of certain translation strategies? How can these translation procedures be classified according to the text genre in which they occur? Is there any type of translation errors which are common to all kind of texts? We will take into account and analyze different parameters of the translation process in order to characterize the most frequent factors and check if there is some observable trend or repeated patterns.

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. Being in its early stages, the focus will be given to the text compilation process and show some of the first data stored in the corpus. The samples of the texts collected comprise a wide variety of genres, ranging from more generic texts to more specialized ones: we have managed to comprise texts from audiovisual translations (including dubbing, subtitling for hearing population and for deaf and hard-of hearing population), scientific, humanistic, literary, economic and legal translation texts. 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, engineers studying EFL and MA students, all of them with different English levels (from B1 to C1); for some of the students, this would be their first experience with translation.

The main goal of this specific study is to gather and comment on the first impressions after carrying out the corpus compilation process and the first attempts at testing the search of any significant data in the texts collected, along with the major purpose of the project which aims at comparing the most frequent translational behaviors of students, as well as identifying and analyzing recurrent strategies and translation errors (including terminological, phraseological, or structural mistakes) in the students' production. The research approach is based on compiling, organizing, aligning and tagging the texts which will be analyzed in a subsequent stage.

The MUST corpus is searchable via the Hypal4MUST, a web-based interface developed by Adam Obrusnik from Masaryk University (Czech Republic), which includes a translation-oriented annotation system. We are currently testing the Hypal4MUST, and we will explain the drawbacks and advantages that we have identified in our first experiences with this interface. A distinctive feature of the interface is that it allows source texts and target texts to be aligned, but we have to check its availability with audiovisual translations.

References

Granger, S. & Lefer, M.-A. (2017). Bridging the gap between learner corpus research and translation studies: The Multilingual Student Translation corpus. *4th Learner Corpus Research Conference*, Bolzano, Italy, 4-7 October 2017.

Johansson, S. (2007). *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Obrusnik, A. (2014). Hypal: A User-Friendly Tool for Automatic Parallel Text Alignment and Error Tagging. *Eleventh International Conference Teaching and Language Corpora*, Lancaster, 20-23 July 2014, 67-69.

Rica, J. P. (in press 2018). *Corpus Studies and Audiovisual Translation: Subtitling.* Series: New Trends in Translation Studies (Edited by J. Díaz Cintas). Frankfurt: Peter Lang.

Rica, J. P., Albarrán, R. & 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.

Rica, J. P. & Braga, J. (2015). *Herramientas y técnicas para la traducción inglés-español: los textos literarios*. Madrid: Escolar y Mayo.

Evaluation of distributional semantic models for the extraction of semantic relations from small specialized corpora

Juan Rojas-Garcia, Pamela Faber
University of Granada
juanrojas@ugr.es, pfaber@ugr.es

Corpus-based lexical studies on specialized domains and for specific purposes normally rely on small corpora, which, in the case of written ones, range from 250,000 to around 6 million tokens (Flowerdew 2004: 19; O'Keeffe et al. 2007: 4). The small size of specialized written corpora is accepted by the scientific community, since some scholars have observed that small corpora are appropriate for studying high-frequency vocabulary (Hunston 2002; Kennedy 1998). In addition, as stated by Meyer (2004), McEnery & Wilson (2001) and O'Keeffe (2007), small corpora are generally better designed and carefully sampled to maximally represent the language phenomena under investigation, so that they can be manually approached. In doing so, even with relatively small amounts of data, O'Keeffe et al. (2007: 198) pointed out that "specialized lexis and structures are likely to occur with more regular patterning and distribution" than in a large, general corpus. Therefore, small specialized corpora have proved to be useful in giving insights into patterns of language use in particular settings (Koester 2010: 67).

With the advent of computers and electronic corpora, scholars of lexical semantics widely use statistical methods such as: (a) association measures applied to Distributional Semantic Models (DSMs) to identify contextual clues in corpus data that are indicative of either the meaning of a term (collocations and colligations) or the semantic relation held between two terms (hyponymy/hypernymy, causality, etc.), and (b) clustering techniques to classify the occurrences of terms into distinct senses (for polysemous terms) or semantic categories based on these contextual clues. Therefore, in this paper we focused on how DSMs, as a statistical state-of-the-art technique, can support the lexicological analysis of semantic relations between terms in a small domain-specific corpus.

More specifically, we concentrated on the relations of hyponymy/hypernymy, meronymy, causality, location and function in a subcorpus of English texts on Coastal Engineering, comprising 6 million tokens and composed of specialized and semi-specialized texts. This subcorpus is integral part of the English EcoLexicon corpus, which currently contains over 59 million words in English and is focused on the environmental domain. It was manually compiled for the development of EcoLexicon (http://ecolexicon.ugr.es), an electronic, multilingual, terminological knowledge base on environmental sciences. To maximize representativeness, the corpus was designed based on criteria proposed by Sinclair (1991, 2005), Meyer (2004), and Biber (2008): balance, diversity of sources, availability of texts in electronic form, period, size, use of complete texts, and variety of writers.

DSMs represent the meaning of a term as a vector by considering the statistics of its co-occurrence with other terms in the corpus. The distributional hypothesis lies at the heart of DSMs, which led to the finding that semantically similar terms tend to have similar contextual distributions (Miller & Charles 1991). In addition, the construction of a suitable DSM for a particular task is highly parameterised, and numerous studies have addressed the evaluation and optimization of DSMs in very large, general corpora (Bullinaria & Levy 2007, 2012; Baroni et al. 2014; Kiela & Clark 2014; Lapesa et al. 2014). However, to the best of our knowledge, the capabilities of DSMs in capturing different semantic relations in small specialized corpora have not been evaluated. As such, the overall aim of this paper was to look for parameter combinations, suitable for each semantic relation, which are both efficient from a statistical point of view, and relevant from a semantic standpoint. For that purpose, an experiment was carried out in which DSMs were built on domain-specific corpora, and then evaluated on gold standard data extracted from EcoLexicon database.

We selected the following set of parameters for investigation:

- **Corpus size:** We used a specialized corpus comprising 6 million tokens as starting point, and split the data into bins of varying sizes. We produced subcorpora ranging from 0.5 to 6 million tokens in steps of 0.5 million tokens. In total, 12 corpora were evaluated.
- **Size of context window:** Contexts are determined by term co-occurrences within a sliding window of a given size, where the window simply spans a number of terms occurring around instances of a target term. We considered window sizes ranging from 1 to 15 terms.
- **Shape of context window:** The shape is a function that determines the increment that is added to the co-occurrence frequency of a given target term-context term pair, based on the distance between both terms. In a rectangular window, this increment is always 1, regardless of distance. In a triangular window, the increment is inversely proportional to the distance between the target term and the context term.
- **Weighting:** Weighting schemes increase the importance of context terms that are more indicative of the meaning of the target term. Co-occurrence counts were weighted using the following association measures: simple log-likelihood, Mutual Information, t-score, z-score and tf.idf (Evert 2005).
- **Weighting transformation:** A transformation function was applied to reduce the skewness of weighting scores, following Lapesa et al. (2014). The transformations were: none, square root, and logarithmic.
- **Dimensionality reduction:** Both linear and non-linear dimensionality reduction techniques were applied to project distributional vectors to a relatively small number of latent dimensions. The most commonly Singular Value Decomposition technique (Golub & Van Loan 1996) was found to achieve better results in small general corpora (Lapesa et al. 2014). Nevertheless, to the best of our knowledge, no evaluation has been carried out to study the impact of different techniques of dimensionality reduction on the performance of DSMs.
- **Type of model:** count models such as *Latent Semantic Analysis* (LSA) (Landauer & Dumais 1997), *Hyperspace Analogue to Language* (HAL) (Lund & Burgess 1996), and *Correlated Occurrence Analogue to Lexical Semantic* (COALS) (Rohde et al. 2006); and predictive models such as *word2vec* (Mikolov et al. 2013a, 2013b), *GloVe* (Pennington et al. 2014), and *Parallel Document Context* (PDC) (Sun et al. 2015) were explored.

Using the gold standard dataset representing five kinds of semantic relations, we evaluated the DSMs and analysed the influence of each model parameters. The results obtained indicate that the model with the best performance depends on the targeted relations, and that the influence of model parameters varies considerably with respect to this factor.

References

Baroni, M., Dinu, G. & Kruszewski, G. (2014). *Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore (Maryland, USA): ACL, 238-247.

Biber, D. (2008). Representativeness in corpus design. In T. Fontenelle (ed.) *Practical lexicography. A reader*. Oxford: Oxford University Press, 63-87.

Bullinaria, J. A. & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39(3), 510-526.

Bullinaria, J. A. & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods* 44(3), 890-907.

Evert, S. (2005). *The Statistics of Word Cooccurrences*. PhD dissertation. Stuttgart: Stuttgart University.

Flowerdew, L. (2004). The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings. In U. Connor & T. Upton (eds). *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, 11-33.

Golub, G. & Van Loan, C. (1996). *Matrix Computations* (3rd ed.). Baltimore: JHU Press.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kiela, D. & Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*. Gothenburg (Sweden): EACL, 21-30.

Koester, A. (2010). Building small specialized corpora. In A. O'Keeffe & M. McCarthy (eds). *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 66-79.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2), 211-240.

Lapesa, G., Evert, S. & Schulte im Walde, S. (2014). Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (SEM 2014)*. Dublin: SEM, 160-170.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers* 28(2), 203-208.

McEnery, A. & Wilson, A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Meyer, C. (2004). *English corpus linguistics. An introduction*. Cambridge: Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G. S. & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations*. Scottsdale (Arizona): ICLR.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Weinberger (eds). *Advances in Neural Information Processing Systems*. Stateline (Nevada, USA): Curran Associates, 26, 3111-3119.

Miller, G. & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1), 1-28.

O'Keeffe, A. (2007). The Pragmatics of Corpus Linguistics. In *Proceedings of the Fourth Corpus Linguistics Conference*. Birmingham: University of Birmingham.

O'Keeffe, A., McCarthy, M. J. & Carter, R. A. (2007). *From Corpus to Classroom*. Cambridge: Cambridge University Press.

Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods for Natural Language Processing*. Doha (Qatar): EMNLP, 1532-1543.

Rohde, D. L. T., Gonnerman, L. M. & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM* 8, 627-633.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (ed.) *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books, 1-16.

Sun, F., Guo, J., Lan, Y., Xu, J. & Cheng, X. (2015). Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations. In *Proceedings of the 53rd Annual Meeting of the ASL and the 7th International Joint Conference on Natural Language Processing*. Beijing (China): ACL, 136-145.

Testing the contrastive application of the n-gram method to typologically different languages: The case of English and Czech children's literature

Denisa Šebestová, Markéta Malá
Charles University
sebestovadenisa@gmail.com, Marketa.Mala@ff.cuni.cz

The study sets out to address two issues raised by previous studies dealing with phraseology and children's literature. The first question is methodological, focussing on "the potential contribution" of an n-gram-based approach to language comparison (Granger 2014). N-grams have proved a useful starting point when comparing languages which are linguistically close, and a rather "challenging" one when dealing with typologically different languages (Čermáková & Chlumská 2017; Hasselgård 2017; Ebeling & Ebeling 2013), such as predominantly analytical English and inflectional Czech, which are compared in this study. We test the advantages and limitations of the n-gram method, considering the possibilities of combining it with other quantitative methods (POS-grams, frequency word/lemma lists) as a first step in a comparative analysis.

The frequency and types of n-grams are highly sensitive to language as well as to register (Biber & Conrad 2009: 6). We examine imaginative fiction written for child and teenage audience as a register delimited primarily by its intended audience but also by its linguistic features, which serve specific communicative functions (Hunt 2005; Thompson & Sealey 2007). The study aims to explore to what extent n-grams can help characterise and point out differences between English and Czech children's fiction.

The study relies on comparable English and Czech corpora of children's fiction: two small corpora of approximately 650,000 words each, and two large corpora of approx. 2,700,000 words each – children's literature sub-corpora of the Czech National Corpus (SYN) and the British National Corpus. For technical reasons, we restrict the queries to 250,000 hits in the large corpora. For the time being (this stage of our study being a preliminary probe), we consider this limitation acceptable, as the large corpora present a unique option to use an otherwise inaccessible dataset containing a wide range of children´s fiction. The two small corpora allowed for a detailed examination, whereas the large ones served to test and verify our findings based on the small corpora, supplementing them by lemma and POS queries.

We extracted 2- to 5-grams (i.e. continuous sequences of 2-5 words excluding punctuation) from the smaller English and Czech corpora (with the minimum range set at 2 texts, and the frequency cut-off point at 50, 10, 5 and 3 tokens, respectively). The numbers of n-grams (types and tokens) above the threshold are consistently higher in English; the difference is statistically significant at p < .001 (Table 1). The ratios suggest a much larger extent of recurrent patterning in analytical English than in Czech, characterized by high morphological variability and free word-order (cf. the Czech 4-grams: *se nedá nic dělat, nedá se nic dělat, nedalo se nic dělat*). The slightly higher type/token ratios in Czech again point to higher variability of Czech as compared to English.

	Czech corpus (655 267 tokens)			English corpus (682 371 tokens)		
	types	tokens	type/token ratio	types	tokens	type/token ratio
2-grams	453	55195	0,82	1331	187083	0,71
3-grams	548	9765	5,6	3083	63949	4,8
4-grams	121	876	13,8	1714	14945	11,5
5-grams	40	147	27,2	1013	4422	22,9

Table 1. Numbers of n-grams extracted form the two small corpora

Another difference between the two corpora consists in the representation of verbs and nouns within the most frequent n-grams. Based on the small corpora the percentage of n-grams comprising a verb appeared higher in Czech than in English. This was verified using the larger tagged corpora: the most frequent 3-5-grams comprise verbs in Czech (e.g. pronoun-verb-preposition-noun, *se vydal na cestu*), while the most frequent English ones include prepositions and nouns (e.g. preposition-determiner-adjective-noun, *for a long time*). This is again in accord with the typological expectations, Czech generally preferring (finite) verbal expression and English being more 'nominal'. The POS observations highlighted not only the importance of verbs for Czech but also their high morphological variability as a potential hindrance to the use of the n-gram approach.

Frequent 3-5-grams identified in the small corpora were classified semantically. Both languages contain n-grams which fall into the categories of time (*for the first time, od rána do večera*), space (*the edge of the, na všechny strany*) and modality (*we´ve got to, zdálo se mu*). These categories seem to be essential for the purposes of narrative fiction (Thompson & Sealey 2007: 21). In addition, the English n-grams contained members of other semantic categories, such as verbs of communication or thinking (*I´ll tell you, I don´t think*). The absence of these verbs from the Czech n-grams was surprising as these verbs are to be expected in fiction. Therefore, we looked into frequency lists of verb lemmas occurring in the Czech corpus. They indeed contained verbs of communication or thinking (*říci – 'say', vědět – 'know'*). However, these verbs were not present in the n-grams due to the morphological diversity of the Czech verb forms (e.g. *říci: a řekl jim, a řekl mu, a řekla mu*). This confirms that to examine Czech, a combination of methods is required, including partial lemmatization and perhaps identification of patterns on the basis of n-grams (Ebeling & Ebeling 2013; Gries 2008).

For Czech, frequent 3-5-grams also include idioms in the traditional (taxonomic) sense (*než bys řekl švec*) as well as phraseological units (*to je dost že jdete*), which were not found in the English material (cf. Altenberg 1998: 105-106).

To conclude, the n-gram method proved to be a useful corpus-driven starting point in a contrastive analysis of large quantities of text. While highlighting typological characteristics of the languages compared, it also pointed to semantic similarities and contrasts within the given genre. Complemented by semantic analysis, n-grams show effectively the basic categories present in a narrative text.

The n-gram method has more limitations in Czech due to the inflectional character of the language. Therefore, a combination of methods seems beneficial for the description of Czech, including frequency lists, partial lemmatization and n-gram based patterns.

References

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

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

Čermáková, A. & Chlumská, L. (2017). Expressing PLACE in children's literature. Testing the limits of the n-gram method in contrastive linguistics. In H. Dirdal & T. Egan (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.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (eds). *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins, 3-25.

Hasselgård, H. (2017). Lexical patterns of place in English and Norwegian. In H. Dirdal & T. Egan (eds). *Cross-linguistic Correspondences: From Lexis to Genre*. Amsterdam: John Benjamins, 97-119.

Hunt, P. (2005). Introduction: The expanding world of Children's Literature Studies. In Peter H. (ed.) *Understanding Children's Literature* (2nd ed.). London: Routledge, 1-14.

Thompson, P & Sealey, A. (2007). Through children's eyes? Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics* 12(1), 1-23.

Sources and tools

The British National Corpus. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available from http://bncweb.lancs.ac.uk (Accessed January 2018).

The Czech National Corpus – SYN (version 6): Institute of the Czech National Corpus, Praha. Available from http://www.korpus.cz (Accessed January 2018).

Anthony, L. (2017). *AntConc* (Version 3.5.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Lithuanian discourse markers in utterance initial position: A glimpse at parallel corpus data

Audronė Šolienė
Vilnius University
audrone.soliene@gmail.com

Research on discourse markers (DMs) has been substantially increasing over the past few decades and resulted in a considerable body of studies (see Fraser 1999; Schiffrin 2008; Urgelles-Coll 2010; Amador-Moreno et al. 2015; Auer & Maschler 2016; Brinton 2017; Fedriani & Sansó 2017, *inter alia*). DMs have also achieved a great deal of scholarly attention in a contrastive perspective (Aijmer & Simon-Vandenbergen 2003; Lewis 2006; Degand 2009; Johansson 2007; Beeching & Detges 2014; Furkó 2014). In Lithuanian, DMs (sometimes referred to as 'discourse particles') have been sporadically analysed in terms of their functional classes (Ambrazas 2006a), lexical sources and categorial status (Holvoet & Pajėdienė 2005) as well as diachronic development (Ambrazas 2006b; Nau & Ostrowski 2010). However, contrastive corpus-based Lithuanian-English studies of DMs drawing on empirical data from parallel corpora are rather innovative and rare.

This cross-linguistic study sets out to describe the quantitative and qualitative distribution of the utterance initial Lithuanian DMs *kad*, *na/nu* and *va*, to determine the translational correspondences (TCs) of the DMs in English, as well as to reveal their functional diversity in terms (inter)subjectivity (Traugott 2010), e.g.:

(1) LT-orig: *– Ką jūs čia kaip susitarę vis apie mano sūnų, ko jūs vis... dar vaikas išgirs...*
*– **Kad** Simas gal nieko nesupranta...*
EN-trans: *"Have you agreed, all of you, to go on about my son? Why, are you always … and the child will end up hearing …"*
*"**But** Simas doesn't understand, or does he?"*

(2) LT-orig*:* *– Gerai, gerai. Sakei. Du kartus pasakei.*
*– **Na** matai, – kiek atlyžo Vaitkus.*
EN-trans: *"Well, all right. You have told me that. Twice."*
*"**Well**, you see," Vaitkus softened a bit.*

(3) LT-orig*:* *– Pone Storosta, **va** jūs vedėt vokietę, o pasakykite, tik dovanokit man seniui, ar buvot laimingas (...).*
EN-trans: *"Mr. Storosta, **look**, you married a German, tell me, forgive this old man his bluntness. Were you two happy?"*

The corpus-based approach adopted in this study helps to reveal patterns and meanings of DMs which would be difficult to pin down by mere introspection. The research method is a quantitative and qualitative contrastive analysis based on the data extracted from a self-compiled bidirectional parallel corpus – ParaCorp$_{\text{EN}\rightarrow\text{LT}\rightarrow\text{EN}}$ (Šolienė 2013). The corpus is designed following the model of the English-Norwegian Parallel Corpus (Johansson 2007). The ParaCorp$_{\text{EN}\rightarrow\text{LT}\rightarrow\text{EN}}$ was compiled from original English fiction texts and their translations into Lithuanian and original Lithuanian fiction texts and their translations into English. The size of the corpus is about 5M words. A reference has also been made to the Corpus of the Contemporary Lithuanian Language (CCLL) (http://donelaitis.vdu.lt), namely the sub-corpus of spoken Lithuanian (447, 396 tokens).

The quantitative results show that the most frequent of the three DMs in question appeared to be *na*. Its normalized frequency (f per 10,000 words) is 3.6 in fiction and 19.6 in spoken register, respectively. A note should be made with regard to its variant *nu*, which is considered to be sub-standard in Lithuanian. It is strikingly frequent in spoken Lithuanian – f = 178.5, in contrast to only 0.16 in fiction. The DM *va* comes second in terms of frequency. The least frequent is *kad*. Its use as a DM in fiction is scarce (f = 0.3 only). The data show that *kad* is predominantly used as a subordinator introducing clauses of concession or cause/reason (f = 2.2) or as an optative marking positive or negative wish (f = 0.7).

The analysis of the translational paradigms of the DMs in question has proved that, due to their extreme multifunctionality, non-propositionality, context-dependence and non-referential (interpersonal and textual) function, they exhibit a wide array of different TCs. The DM *na* is the most functionally and translationally versatile: it has an extensive list of TCs, a total of 21 different variants. The wide range of translations of *na* revealed its diverse functions in discourse in terms of (inter)subjectivity: it can be used as a reluctant, reserved or positive response or agreement (*aha*, *uh-uh*, *yes*, *fine*, *okey*, etc.); it may invite the interlocutor to add information and thus hasten the flow of discourse (*ok*, *go ahead*, *well*, etc.); it can indicate impatience (*come on*, *go on*, *ok*, etc.); it may serve as a dialogue opener or express amazement and mirativity (exclamatory structures with *what*). Finally, *na* can occur in passages reproducing inner thought (see Sawicky 2012). The utterance initial *va* resembles *na* in almost most of its functions and the variety of TCs, the only difference being the demonstrative nature of *va*. In such cases it is translated as *here*, *there*, and *look*. *Kad* as a DM usually conveys the inability of the speaker to answer the question and expresses his/her reservation, withdrawal or scepticism towards the content of the previous turn. In these cases it is translated by the connector *but*. However, it seems that *kad,* as a DM (in addition to the functions singled out by Sawicky 2012), can also perform the function of exemplification, which is clearly supported by its English TCs – *for example* or *take for example*.

The last point to note is that the category of zero translation appeared to be unifying for all the DMs under investigation. Aijmer & Altenberg (2001: 29) explain the notion of zero correspondence in translation as a phenomenon when "[e]xpressions that do not contribute to the propositional content are often untranslatable in the sense that an exact equivalent cannot be found in another language". Moreover, zero correspondence may be a result of language-specific conventions or different degree of grammaticalization of DMs (Aijmer & Altenberg 2001: 32).

References

Aijmer, K. & Altenberg, B. (2001). Zero Translations and Cross-Linguistic Equivalence: Evidence from the English-Swedish Parallel Corpus. In A. Hasselgren & L. E. Breivik (eds). *From the Colt´s Mouth, and Other Places: Studies in Honour of Anna-Brita Stenstrőm*. Amsterdam: Rodopi, 19-41.

Aijmer, K. & Simon-Vandenbergen, A. M. (2003). The discourse particle *well* and its equivalents in Swedish and Dutch. *Linguistics* 41(6), 1123-1161.

Amador-Moreno, C. P., McCarffery, K. & Vaughan, E. (2015). *Pragmatic Markers in Irish English.* Amsterdam & Philadelphia: Benjamins.

Ambrazas, V. (ed.) (2006a). *Dabartinės lietuvių kalbos gramatika*. Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Ambrazas, V. (2006b). *Lietuvių kalbos istorinė sintaksė.* Vilnius: Lietuvių kalbos institutas.

Auer, P. & Maschler, Y. (eds).. (2016). *NU / NÅ– A Family of Discourse Markers Across the Languages of Europe and Beyond*. Berlin & Boston: Mouton de Gruyter.

Beeching, K. & Detges, U. (eds). (2014). *Functions at the left and right periphery: Crosslinguistic investigations of language change*. Leiden: Brill.

Brinton, L. J. (2017). *The Evolution of Pragmatic Markers in English. Pathways of Change.* Cambridge: Cambridge University Press.

Degand, L. (2009). On describing polysemous discourse markers. What does translation add to the picture? In S. Slembrouck, M. Taverniers & M. Van Herreweghe (eds). *From will to well. Studies in Linguistics offered to Anne-Marie Simon-Vandenbergen*. Gent: Academia Press, 173-183.

Fedriani, C. & Sansó, A. (2017). *Pragmatic Markers, Discourse Markers and Modal Particles. New perspectives.* Amsterdam & Philadelphia: Benjamins.

Fraser, B. (1999). What Are Discourse Markers? *Journal of Pragmatics* 31, 931-952.

Furkó, B. P. (2014). Perspectives on the Translation of Discourse Markers. *Acta Universitatis Spientiae, Philologica* 6, 181-196.

Holvoet, A. & Pajėdienė, J. (2005). Aplinkybės ir jų tipai. In A. Holvoet & R. Mikulskas (eds). *Gramatinių funkcijų tyrimai*. Vilnius: Lietuvių kalbos institutas, 93-116.

Johansson, S. (2007). *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies.* Amsterdam & Philadelphia: Benjamins.

Lewis, D. (2006). Contrastive analysis of adversative relational markers, using comparable corpora. In K. Aijmer & A. M. Simon-Vandenbergen (eds). *Pragmatic Markers in Contrast*. Oxford: Elsevier, 139-153.

Nau, N. & Ostrowski, N. (2010). Background and perspectives for the study of particles and connectives in Baltic languages. In N. Nau & N. Ostrowski (eds). *Particles and Connectives in Baltic*. Vilnius: Vilniaus Universitetas & Asociacija "Academia Salensis", 1-37.

Sawicky, L. (2012). Responsive discourse particles in Lithuanian dialog. *Baltic Linguistics* 3, 151-175.

Schiffrin, D. (2008). Discourse Markers: Language, Meaning, Context. In D. Schiffrin, D. Tannen & H. Ehernberger Hamilton (eds). *The Handbook of Discourse Analysis*. MA: John Wiley & Sons, 54-75.

Šolienė, A. (2013). *Episteminio modalumo ekvivalentiškumo parametrai anglų ir lietuvių kalbose* (ms.). Humanitarinių mokslų daktaro disertacija. Vilnius: Vilnius University.

Traugott, E. C. (2010). (Inter)subjectivity and (inter)subjectification: A reassessment. In K. Davidse, L. Vandelanotte & H. Cuyckens (eds). *Subjectification, Intersubjectification and Grammaticalization*. Berlin & New York: De Gruyter Mouton, 29-74.

Urgelles-Coll, M. (2010). *The Syntax and Semantics of Discourse Markers: Continuum Studies in Theoretical Linguistics*. London: Continuum.

Syntactic complexity as a stylistic feature of subtitles

Andy Stauder
University of Innsbruck
andy.stauder@uibk.ac.at

Linguistic style has been a popular topic for a while, although a widely adopted definition of the concept is lacking. There seems to be a notion of *choice of linguistic expression*. The concept is similar to that of *translation quality* – quality literally meaning "how-ness", which is similarly poorly defined and thus has elicited what House (1997: 1) calls "anecdotal, biographical and neohermeneutic approaches to judging translation quality", i.e. approaches that have hardly anything to say about linguistic features in the narrow sense of the word.

A more objective and much less vague approach is the seminal corpus-based methodology by Baker (2000). This proposes a clearly defined set of measurable features for describing *translator style*: frequencies of certain words or parts of speech, mean sentence length, standardised type-token ratio, etc., "typical of a [given] translator" (ib.: 245). From this, it is clear that style can be layered: the style of the original may be changed in a consistent fashion by the translator. Also, style does not necessarily have to consist of conscious choices, but may be habitual and due to a variety of influencing factors. These can be described very well with the *diasystem* by Coşeriu (cf. 1981[1958]; Goossens 1977; Faust 1988), which classifies linguistic modes of expression according to social context. Thus, a person's mode of linguistic expression may be influenced by class, age, geographical provenance, historical period, and target audience/situation. All of these may also be fictitious: an author may *want* to have one of their characters sound a certain way. This adds a third layer of style: on top of the personal one of the author, which may be influenced by the aforementioned *dia* factors, comes the one of the possible characters created by the author, which may then be superseded by the alterations due to the translator's style, which may again be influenced by the *dia* factors and conscious choices of the translator.

So, on the one hand, style characterises the way a person writes and translates, and, on the other, can also be used as a creative device for writers (or translators) to shape the characters in their works. This is especially true for audiovisual entertainment: this is usually very character-heavy and one main feature characterising the protagonists is the way they talk. The goal of this research is therefore the following: it is interested in operationalising one feature of linguistic style – syntactic complexity – for Audiovisual Translation, with the results having possible application in the identification of translator style and also translation quality, at least as far as similarity of source and target text are concerned. Thus, the paper's aim could be said to fall within the area of *translation stylometry*. The research is to be conducted on subtitles because these present audiovisual language data in a form that lends itself to machine-processing.

There are few studies that specifically target syntactic complexity as a stylistic feature: most seem to focus on lexical, i.e. semantic and pragmatic (cf. e.g. Kenny 2001; Winters 2004; Saldanha 2011), but not so much syntactic phenomena. Those that do seem to do so indirectly, by examining the feature of readability, and by applying it to whole texts rather than individual sentences/utterances (cf. e.g. the insightful Huang 2015: 95 ff). According to Huang (ibid: 115) "It is found that statistics about readability provided by manual calculation or computer software cannot effectively differentiate one text from another in terms of style." This study, on the other hand, looks to pin down syntactic complexity as a stylistic feature of individual utterances, in the context of the dialogue-heavy field of Audiovisual Translation. Here, as in any (quasi-)literary work, it is not only important what is being said, but also the way it is being said (cf. what Jakobson 1972[1960] called the *secondary structure* of literary texts). Therefore, the research questions are:

A) Is syntactic complexity a significant stylistic feature of linguistic utterances in the form of subtitles?
B) In how far is it possible to distinguish the characters of a TV show using this feature?
C) In how far is the syntactic complexity of a specific TV show's characters' utterances reproduced in their translation?

While finding answers to these research questions, the study also aims to tackle a problem that may not be relevant to readability studies, but which is to translation studies: the scores of linguistic complexity calculated for the subtitles from the test corpus are to be standardised for each language with the help of representative corpora, because a specific English sentence may, in comparison to English in general, be more complex than its German translation in comparison to German in general.

The methodology for measuring syntactic complexity is one devised by the author and takes into account word count per sentence, syllable count per word, and dependency tree-path depth. The reason for devising a dedicated methodology is that classical readability or syntactic complexity calculation methods are somewhat limited and may not capture the phenomenon of syntactic complexity adequately: they mostly limit themselves to quantitative features (word and sentence length; cf. e.g. Flesch 1948; Björnsson 1968) or account for structural information only heuristically, e.g. by including verb count (cf. Fichtner 1981). The proposed methodology is to consist in clustering subtitles (using the Stylo R package: Eder et al. 2016) according to their syntactic complexity and trying to assign each cluster to a character from the respective show. It is then to be examined in how far this clustering differs from one performed with the subtitles' translations.

The test material is to consist of a parallel corpus compiled by the author, containing the English and German subtitles of the show *Two and a Half Men*. The corpora for standardising the scores are from the Aranea family (cf. Benko 2014).

Possible limitations might be the fact that subtitling tends to generally simplify syntax and the fact that there may be several characters on a show exhibiting similar levels of syntactic complexity. The scope of these effects may also be discussed in the study.

References

Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator? *Target* 12(2), 241-266.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček & K. Pala (eds). *Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655.* Springer International Publishing Switzerland, 2014, 257-264.

Björnsson, C. H. (1968). *Läsbarhet* (with an English summary). Stockholm: Bokförlaget Liber.

Björnsson, C. H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly* 18(2), 480-497.

Coşeriu, E. (1981 [1958, oral presentation]). Los conceptos de 'dialecto', 'nivel' y 'estilo de lengua' y el sentido propio de la dialectologia. *Lingüística española actual III*, 1-23. [n.v.]

Eder, M., Rybicki, J. & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal* 8(1), 107-121, https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Fichtner, E. G. (1981). Measuring Syntactic Complexity: The Quantification of One Factor in Linguistic Difficulty. *Die Unterrichtspraxis / Teaching German* 13(1), 67-75.

Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology* 32(3), 221-233.

House, J. (1977). *A Model for Translation Quality Assessment*. Tübingen: Narr.

Huang, L. (2015). *Style in Translation: a Corpus-Based Perspective*. Berlin: Springer.

Kenny, D. (2001). *Lexis and Creativity in Translation. A Corpus-based Study.* Manchester: St. Jerome.

Saldanha, G. (2011). Style of Translation: The Use of Source Language Words in Translations by Margaret Jull Costa and Peter Bush. In A. Kruger, K. Wallmach & J. Munday (eds). *Corpus Based Translation Studies: Research and Applications*. New York: Continuum, 237-258.

Winters, M. (2004). F. Scott Fitzgerald's Die Schönen und Verdammten: A corpus-based study of loan words and code switches as features of translators' style. *Language Matters, Studies in the Languages of Africa* 35(1), 248-258.

The Obama presidency, *the Macintosh keyboard* and *the Norway fiasco*: English proper noun modifiers in German and Swedish contrast

Jenny Ström Herold, Magnus Levin
Linnaeus University
jenny.strom.herold@lnu.se, magnus.levin@lnu.se

Nouns used as premodifiers have tripled over the last two centuries in English (Biber et al. 2009: 187), and proper nouns are increasing in frequency in writing, a change which is particularly noticeable with acronyms (Leech et al. 2009: 212). In German and Swedish, which disallow nouns as premodifiers (**Dylan bootlegs*; **Australien Projekt*) and instead use either hyphenated or solid compounds (*Dylan-bootlegs* (Sw.); *Australienprojekt* (Ge.)), the frequencies of such compounds also appear to be on the increase (for German, see Zifonun 2010 and for Swedish Koptjevskaja-Tamm 2013). It is noteworthy that Zifonun (2010) attributes this change in German to English influence.

Although previous studies of English proper noun modifiers have touched upon contrastive aspects (see e.g. Koptjevskaja-Tamm 2013; Schlücker 2013: 464-465; Breban 2017: 13), there has to date been no systematic study. The aim of this paper is to fill this gap by investigating the semantic categories personal names, place names and names of organizations used as premodifiers in both English source texts and English target texts translated from German and Swedish. The investigation shows (i) what structural means are used in German and Swedish to render the modifiers, (ii) in what ways the semantic categories of the proper nouns affect the translation choices, (iii) what German and Swedish structures are translated as English proper noun modifiers and (iv) the specific nature of translated language (cf. Baker 1993).

The corpus used in this study, the Linnaeus University English-German-Swedish Corpus (LEGS) (see Ström Herold & Levin forthcoming), consists of recently published (2000s) popular non-fiction texts (e.g. biographies and popular science) in English, German and Swedish, and is balanced for the three languages, each original always being accompanied by two target texts. Also, each author and translator is represented only once. The corpus, which is being compiled by the present authors, currently comprises about 250,000 words in each source language with translations. The main advantage of the corpus is that there are always two translations available for every source-text segment. This makes it possible to compare how the very same instance has been translated into two target languages, thereby allowing identification of language-specific and translation-specific features. Moreover, the corpus provides translations from two source languages into each language. A tagged version of the corpus was searched for proper nouns immediately followed by (a) common noun(s). This way, more than 1,000 instances of English proper noun modifiers and 1,600 German and Swedish correspondences were retrieved.

The results show that there are many different alternatives among the renderings of proper noun modifiers, the three most frequent being compound nouns (*the Norway fiasco* > *das Norwegen-Fiasko* (Ge.)), prepositional phrases (*the Apple corridors* > *korridorerna på Apple* (Sw.)) and genitives (*U.N. climate summits* > *FN:s klimattoppmöten* (Sw.)). Apart from these, ten minor correspondence categories were identified.

Among the notable language-specific tendencies is a significantly stronger German preference for compounds (*the Stanford campus* > *den Stanford-Campus*) (cf. Carlsson's 2004 finding on compounds being more common in German than in Swedish). Swedish translations instead use more postmodifying prepositional phrases (*the Fukushima disaster* > *katastrofen i Fukushima* ['the disaster in Fukushima']). However, compounds are strongly disfavoured in both German and Swedish translations when the noun phrase contains a "heavy head" (cf. Koptjevskaja-Tamm 2013), i.e. a head consisting of a 'compound'/ two or more nouns. Such noun phrases are instead often translated into (compound nouns followed by) prepositional phrases containing the proper nouns, e.g. *a Yale law degree* > *einen Juraabschluss in Yale* (Ge.); *juristexamen vid Yale* (Sw.) ['a law-degree at Yale'].

Concerning the semantic categories of proper noun, the ones based on organizations are typically translated into compounds (*every Apple product* > *jedes Apple-Produkt* (Ge.)) or genitives (*Red Army soldiers* > *Röda arméns soldater* (Sw.)). In contrast, proper noun modifiers based on place names are more often rendered as prepositional phrases (*Ontario residents* > *die Bürger von Ontario* (Ge.); *invånarna i Ontario* (Sw.)), as already noted by Schlücker (2013) for German.

Overall, acronyms are quite frequent as premodifiers (*NKVD troops*) in both English source texts and translations, and they have a bearing on translation choices. While German prefers compounds (*a US news show* > *einer US-Nachrichtensendung*), Swedish prefers genitives (*U.S. negotiators* > *USA:s förhandlare*).

Most of the proper noun modifiers in English target texts translated from German and Swedish are based on compounds (e.g. *DDR-Fernsehen* (Ge.) > *GDR television*). Postmodifying prepositional phrases are very rarely translated into premodifiers (*ett hotell i Florida* (Sw.) > *a Florida hotel*), as also found by Levin & Ström Herold (2017), and the same holds true for genitives. It is noteworthy that some English modifiers originate in the translation strategy explicitation (*skärgården* ['the archipelago'] (Sw.) > *the Stockholm archipelago*).

The results indicate that premodifiers (such as proper noun modifiers) are rarer in translations than in source texts, possibly because they are less explicit and/or more compressed, as suggested by Levin & Ström Herold (2017). Another translation-specific feature concerns proper noun modifiers being dispreferred with unknown/exotic elements, as when the Swedish compound *Expressenjournalisten* is translated into a postmodifying prepositional phrase in English *a journalist on Expressen newspaper*, in spite of similar constructions often being written as premodifiers in English source texts (e.g. *the Time reporter*).

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.

Biber, D., Grieve, J. & Iberri-Shea, G. (2009). Noun phrase modification. In G. Rohdenburg & Julia Schlüter (eds). *One Language, Two Grammars? Differences between British and American English*. Cambridge: Cambridge University Press, 182-193.

Breban, T. (2017). Proper names used as modifiers: a comprehensive functional analysis. *English Language and Linguistics* 1-21.

Carlsson, M. (2004). *Deutsch und Schwedisch im Kontrast: Zur Distribution nominaler und verbaler Ausdrucksweise in Zeitungstexten.* Göteborg: Acta Universitatis Gothoburgensis.

Koptjevskaja-Tamm, M. (2013). A Mozart sonata and the Palme murder: The structure and uses of proper-name compounds in Swedish. In K. Börjars, D. Denison & A. Scott (eds). *Morphosyntactic Categories and the Expression of Possession*. Amsterdam & Philadelphia: Benjamins, 253-290.

Leech, G., Hundt, M. Mair, C. & Smith, N. (2009). *Change in Contemporary English. A Grammatical Study*. Cambridge: Cambridge University Press.

Levin, M. & Ström Herold, J. (2017). Premodification in translation English hyphenated premodifiers in fiction and their translations into German and Swedish. In T. Egan & H. Dirdal (eds). *Cross-linguistic Correspondences: From Lexis to Genre*. Amsterdam & Philadelphia: Benjamins, 149-176.

Rosenbach, A. (2007). Emerging variation: determiner genitives and noun modifiers in English. *English Language and Linguistics* 11(1), 143-189.

Schlücker, B. (2013). Non-classifying compounds in German. *Folia Linguistica* 47, 449-480.

Ström Herold, J. & Levin, M. (forthcoming). English *ing*-clauses and their German and Swedish correspondences.

Zifonun, G. (2010). Von *Bush administration* zu *Kohl-Regierung*: Englische Einflüsse auf deutsche Nominalkonstruktionen? In C. Scherer & A. Holler (eds). *Strategien der Integration und Isolation nicht-nativer Einheiten und Strukturen*. Berlin: De Gruyter, 165-182.

Edited vs. translated English: Normalisation in mediated language through the lens of the split infinitive

Helen Swallow
Université catholique de Louvain
helen.swallow@uclouvain.be

In 1993 Baker posited the theory of translation 'universals', i.e. features considered to be characteristic of all translated texts. These are also referred to as 'translationese' (Baker 1993; Rayson *et al.* 2008), the 'third code' (Frawley 1984; Granger in press) or 'recurrent features of translated language' (Redelinghuys & Kruger 2015). It has since been suggested by Ulrych and Murphy (2008) and others that these universals may be in play in other types of mediated language, including edited language. For example, Kruger (2017: 146) notes that 'there is […] considerable […] support for the hypothesis that editors demonstrate a tendency towards conventionalisation or normalisation'. This study focuses on the normalisation/conservatism feature, defined by Baker (1996: 176-177) as '[t]he tendency to conform to patterns and practices which are typical of the target language, *even to the point of exaggerating them*' [my italics].

While much research has been dedicated to universals in translation, relatively little work has been done on this feature in edited language. Murphy (2008: 182) points out that 'research into edited-mediated language in the EurComm corpus is still in its infancy', and notes that further research is needed. In 2012 Kruger carried out a corpus-based study of the mediation effect in translated and edited language. She concluded that 'this study has produced almost no evidence for a mediation effect that is shared by translated and edited language' (Kruger 2012: 282), adding that 'editing may have its own, peculiar effects' (ibid: 383) and 'involve a different kind of mediation altogether'. Building on Kruger's work, Bisiada (2017: 264) finds, with reference to normalisation, that 'the editing process does not significantly change the features of the language of translation'.

The need for further investigation of mediation effects in edited and translated language has prompted the corpus-based project of which this presentation gives a general overview and a specific illustration in the form of a case study of the split infinitive.

In terms of the corpora used, particular emphasis is placed on the corpus of edited texts, as this corpus type is still relatively rare, and a new corpus of edited English (European ParlIament Corpus of Edited Texts – EPICET) is currently under construction by the author. EPICET collects unedited and edited parliamentary texts (the resolution sections of draft own-initiative reports and opinions) produced in English in the European Parliament (EP) between 2012 and 2016. The original texts are drafted by parliamentary committee officials who are native speakers of any of the EU official languages (looked upon for the purposes of this study as advanced non-native users of English); the texts are subsequently edited by members of the EP translation directorate-general's Editing Unit, all of them English native speakers (the term here used in its commonly accepted sense, while acknowledging that it is the subject of discussion (e.g. Davies 2003)). The originality of EPICET lies in its potential to help remedy the current paucity of dedicated corpora of edited advanced non-native English.

In addition to EPICET, an existing Europarl corpus of texts translated into English will be used, pending the creation by the author by 2019 of a corpus of Europarl translated English texts of exactly the same type as those constituting EPICET, ensuring full comparability between the mediated corpora. Other reference corpora of native-speaker English will also be used, such as the 'newspaper' subcorpus of the new British National Corpus.

To illustrate the project's approach, a case study of the split infinitive (*to* + adverb + infinitive*,* as in *to fully understand*) in edited, translated and unmediated native-speaker English texts will be presented. Madrassa (2009: 99) notes that 'Whilst all modern grammarians agree that the split infinitive […] is not a grammatical error, […] the split infinitive […] remains highly contentious' (ibid: 130). Jang & Choi (2014: 56) state that 'Even though the use of split infinitives has been drastically increased in contemporary English, the issue of whether split infinitives are grammatically acceptable or not does not still seem to be resolved'. Strict adherence to the convention that infinitives should not be split can be seen as a facet of conservatism, and our aim in this case study is to ascertain the extent to which mediated language (edited and translated) adheres to the convention, and how they compare with each other in this respect.

The main research question explored in the study is the following: Do editors and/or translators exaggerate, in comparison with the unmediated native speaker reference texts, the practice of keeping infinitives unsplit, in line with Baker's normalisation / conservatism universal (Baker 1996)? This will be investigated by an examination of the split/unsplit infinitive ratio in edited, translated and unmediated texts.

Within this process, the following subquestions will be answered:

(a) Does a comparison of the edited and unedited texts show an editor bias towards correcting or not correcting split infinitives? To what extent, when producing new text, do editors introduce split/unsplit infinitives of their own?

(b) To what extent does a comparison of the split/unsplit ratio in the unmediated reference texts with those established in the edited and translated corpora substantiate the presence of the 'conservatism' universal in mediated language? Is there a significant difference between the two types of mediated language from this point of view?

Factors which may have influenced editors' and translators' choices will be considered, including style guides and internal editorial guidelines, as will the influence of the lexis, whereby, as noted by Jang & Choi (2014: 65), 'only a limited number of adverbs can split the infinitives and only a number of verbs can be split by these adverbs'.

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: John Benjamins, 233-250.

Baker, M. (1996). Corpus-based Translation Studies: An Overview and some Suggestions for Future Research. In H. Somers (ed.) *Terminology, LSP and Translation Studies in Language Engineering, in Honour of Juan C. Sager*. Amsterdam: John Benjamins, 175-186.

Bisiada, M. (2017). Universals of editing and translation. In S. Hansen-Schirra, O. Czulo & S. Hofmann (eds). *Empirical Modelling of Translation and Interpreting*. Berlin: Language Science Press, 241-275.

Davies, A. (2003). *The Native Speaker: Myth and Reality.* Clevedon: Multilingual Matters.

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (ed.) *Translation: Literary, Linguistic, and Philosophical Perspectives*. Newark: University of Delaware Press, 159-175.

Granger, S. (in press). Tracking the third code: A cross-linguistic corpus-driven approach to metadiscursive markers. In A. Cermakova & M. Mahlberg (eds). *Corpus as Discourse*. Amsterdam: Benjamins.

Jang, Y. & Choi, S. (2014). Split infinitives in English: A corpus-based investigation. *Linguistic Research* 31(1), 53-68.

Kruger, H. (2012). A corpus-based study of the mediation effect in translated and edited language. *Target* 24, 355-388.

Kruger, H. (2017). The effects of editorial intervention: implications for studies of the features of translated language. In G. de Sutter, M-A. Lefer & I. Delaere (eds). *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin: De Gruyter, 113-156.

Mitrasca, M. (2009). The Split Infinitive in Electronic Corpora: Should There Be a Rule? *Concordia Working Papers in Applied Linguistics*, 2.

Murphy, A. (2008). *Editing Specialized Texts in English.* Milan: LED.

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, 293-325.

Ulrych, M. & Murphy, A. (2008). Descriptive Translation Studies and the Use of Corpora: Investigating Mediation Universals. In C. Taylor Torsello, K. Ackerley & E. Castello (eds). *Corpora for University Language Teachers*. Bern: Peter Lang, 141-166.

Not everyone enjoys being loved, but I like it: A contrastive study of three feeling-denoting verbs in English and Norwegian

Øyvind Thormodsæter
University of Oslo
oyvind.thormodsater@ilos.uio.no

Idiomatic use of verbs denoting feelings requires a high command of a language, since such verbs are liable to subjective interpretation both in terms of their inherent intensity and their acceptability in context, and also because their usage to a large extent will be governed by cultural conventions. As such, verbs denoting feelings are interesting from a phraseological point of view, in which context-appropriate multi-word units are thought to be at the core of language production.

This presentation reports on a segment of a larger study of the phraseology of the three English-Norwegian verb pairs ENJOY-NYTE, LOVE-ELSKE and LIKE-LIKE. The study seeks to find out what correspondence patterns and translation paradigms may reveal about similarities and differences between the lexemes in the two languages, how similar meaning is conveyed in the two languages, and how English native speakers and Norwegian learners of English use and understand the English lexemes similarly and differently in context. LOVE, LIKE and ENJOY fit the criteria mentioned: they have overlapping meanings and connotations, their usage will be governed partly by cultural conventions, and they intuitively belong to different parts of the emotional "intensity scale", which makes them eligible for analysis both from an L1 and an L1-L2 perspective. The Norwegian verbs ELSKE, NYTE and LIKE were chosen for comparison because they are considered the closest equivalents of the respective English verbs based on listings in a number of authoritative bilingual dictionaries.

In this investigation, recurrent sequences including the lexemes will be investigated to determine their selectional preferences in original and translated texts in both languages with the intention of mapping cross-linguistic similarities and differences in use and lexicogrammatical features. The investigation will primarily be corpus-based, and the methodology used is largely based on Gilquin's (2000/2001: 98-101) modified version of Granger's (1994) Integrated Contrastive Model, in which Contrastive Analysis (CA) between languages is combined with Contrastive Interlanguage Analysis (CIA). In addition to a qualitative analysis of a selection of examples and a lexicogrammatical categorisation of the material inspired by Hunston & Francis' (2000) pattern grammar, Altenberg's (1999: 254) formula for calculating mutual correspondence (MC) will be used to indicate the level of correspondence between the lexemes in the bidirectional CA part of the study. Data will be extracted from various corpora for the different parts of the investigation, including the ENPC[1] and the written part of the BNC[2] and LBK[3] for the Contrastive Analysis and BAWE[4], MICUSP[5], ICLE[6] and a self-compiled corpus of Norwegian upper-secondary student texts in English and Norwegian for the CIA analysis. This data will finally be compared to data from an elicitation test in order to address the following research questions:

1. What can an analysis of the lexicogrammatical features and translation paradigms of three English-Norwegian verb pairs denoting feeling tell us about:
    a. The level of correspondence between the English and the Norwegian verbs in terms of meaning, usage and selectional preferences/collocations?
    b. The level of consistency in meaning in the individual lexicogrammatical pattern for each verb?

[1] English-Norwegian Parallel Corpus
[2] British National Corpus (English source texts)
[3] Leksikografisk Bokmålskorpus (Norwegian source texts)
[4] British Academic Written English corpus
[5] Michigan Corpus of Upper-Level Student Texts
[6] The International Corpus of Learner English

2.

a. Is there a systematic difference in how English native speakers and Norwegian learners use and understand these and semantically related verbs?
b. To what extent can the differences in usage be seen as a result of lexicogrammatical differences between the two languages?

The segment presented here will draw on data from the ENPC, and will address questions 1a and 1b. Preliminary searches indicate some clear differences between e.g. ENJOY and NYTE, both in terms of semantic preferences, semantic scope and structural features. In brief, ENJOY is more versatile than NYTE both semantically and syntactically, and NYTE more consistently expresses a strong emotion or intensity. Several patterns have a relatively consistent meaning in the examples found, particularly verb + pronouns and verb + non-finite clauses. Examples of ENJOY being translated into ELSKE [LOVE] and LIKE [LIKE] and of NYTE being translated into LOVE and LIKE indicate an overlap in meaning between the lexemes across languages, as exemplified in (1)-(3) below:

1. (…) it occurred to him that she was just the sort of woman who **would enjoy** ten minutes' sex while changing for dinner, (…). (FW1)
   (…) det slo ham at hun var akkurat den type kvinne som **ville nyte [enjoy]** ti minutters sex mens hun skiftet til middag, (…). (FW1T)

2. She **did**, however, **enjoy** the people sitting around and talking, the sociable atmosphere, (…). (DL1)
   Men hun **likte [liked]** at folk satt rundt og pratet, den selskapelige atmosfæren, (…). (DL1T)

3. (…) he **enjoyed** watching the way his canvases drank up black (…). (JH1)
   (…) han **elsket [loved]** å se hvordan lerretene suget til seg sort (…). (JH1T)

The preliminary analysis of ENJOY also lends support to Sinclair's (1999: 158) claim that words often do not express what is considered their "core meaning", and that "few [words] have a clear meaning independent of the cotext" (ibid). Also, the initial analysis shows a relatively low mutual correspondence of 25 % between ENJOY and NYTE, which indicates that there are clear areas of contrast across the two languages. These findings largely concur with Johansson's (2007) findings about LOVE and HATE and their Norwegian counterparts ELSKE and HATE, indicating that some contrastive points are relevant for more than the individual verb pair.

References

Altenberg, B. (1999). Adverbial connectors in English and Swedish: A corpus-based contrastive study. In H. Hasselgård & S. Oksefjell (eds). *Out of corpora: Studies in honour of Stig Johansson*. Amsterdam: Rodopi, 249-268.

Gilquin, G. (2000/1). The ICM: Spicing up your data. *Languages in Contrast* 3(1), 95-123.

Granger, S. (1994). From CA to CIA and back. In K. Aijmer, B. Altenberg & M. Johansson (eds). *Languages in Contrast: Papers from a Symposium on Cross-linguistic Studies*, 37-52.

Hunston, S. & Francis, G. (2000). *Pattern grammar: a corpus-driven approach to the lexical grammar of English.* Studies in Corpus Linguistics (Vol. 4). Amsterdam: John Benjamins.

Johansson, S. (2007). Loving and hating in English and Norwegian: A corpus-based contrastive study. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*, 95-105.

Sinclair, J. (1999). A way with common words. In H. Hasselgård & S. Oksefjell (eds). *Out of corpora: Studies in honour of Stig Johansson,* 157-179.

Quelle traduction! A study of translation of the English exclamative clause into French

Faye Troughton
University of Mons
faye.troughton@umons.ac.be

This paper reports on a study into how the English exclamative clause using *what* is translated into French. In doing so, the paper aims to shed light on cross-linguistic similarities of *wh*- exclamatives, their use in practice, and the translation of degree modification.

The exclamative clause type in English is defined in numerous reference grammars as being formed using the interrogative words *what* (1) and *how* (2) (Quirk et al. 1972: 406; Huddleston 1984; Biber 1999). It is also defined by word order, the potential absence of subject-auxiliary inversion distinguishing it from the interrogative clause (3), while exclamative clauses with *what* are further distinguished by the presence of the indefinite article. It is important to highlight that in this paper the term exclamative refers solely to this "category of form rather than meaning" (Trotta 2000: 101). Other potential syntactic realisations of exclamation, such as inversion constructions (4), nominal constructions with a degree reading (5), and sentence exclamations (7), to name but a few examples, are not addressed (Elliot 1974; Michaelis & Lambrecht 1996; Huddleston & Pullum 2002: 923-924; Zanutti & Portner 2003; Collins 2005; Siemund 2015: 698). The focus of this paper will further be refined by restricting analysis to the exclamative clause using *what (a)* (1) and how it is translated into French.

(1) *What a mess we're in!* (Quirk et al. 1974: 407)
(2) *How delightful her manners are!* (Quirk et al. 1974: 406)
(3) *How delightful are her manners?*
(4) *Quel bruit les manifestants faisaient!* (Jones 1996: 519)
(5) *Did you make a mess!*
(6) *The mess we're in!*
(7) *You made a mess!*

Cross-linguistically, interrogative words, such as *what*, are said to characterise exclamatives (Michaelis 2001: 1042) and accordingly, a number of French reference grammars highlight *quel* as both interrogative and exclamative (Grevisse & Goosse 2008: 505; Riegal et al. 2009: 688). Furthermore, it has been claimed that French exclamative constructions fronted by *quel* (2) correspond, at least in a grammatical sense, with the English exclamative fronted by *what a* (Jones 1996: 519). Indeed, structurally speaking both constructions show similarities. Both are distinguished from the interrogative by word order, in that they do not require subject-auxiliary inversion, and both also exist in ellipted forms (7-8). These observations however, have thus far remained largely theoretical.

(7) *What a mess!*
(8) *Quel bruit!*

Thus, the principle aim of this study is to investigate whether the exclamatives formed using *what* and *quel* accepted in English and French can be argued to be equivalent constructions in terms of their use in translation from English to French.

This will also allow for an investigation into the translation of degree modification in terms of the *what* exclamative. Bolinger (1972) included exclamative *what* in his discussion of "degree words", claiming the "tell-tale indefinite article" to be a signal of *what*'s degree modifying function (Bolinger 1972: 60). Much recent work into the English exclamative clause has also focused on it as an expression of extreme degree (Rett 2008, 2011; Siemund 2017). To illustrate this, construction (7), *what a mess*, expresses the extreme degree of mess as perceived by the speaker. This may equally be applied to any modifiers that precede the noun. When acting upon a non-degree noun, *what* modifies the degree of an unspoken but contextually implied characteristic of this noun, be this qualitative or quantitative. In (9) for example, *what* may modify the degree to which his end was long, admirable, unexpected, or painful, depending on the given context.

(9) *What an end to a man!*
(10) *What an end to a man who made his name as a CND and peace campaigner!* (Europarl)

While as much has also been said of the French *quel* in some grammars (Grevisse & Goosse 1990: 506; Jones 1996: 519), more recently Marandin has classified *quel* as a “mot exclamatif non-scalaire” (2010: 39). He claims that whether a degree interpretation of *quel* is available depends on the presence of a degree or non-degree noun, and in the latter case the exclamative using *quel* expresses “l’idéal ou l’anti-idéal de la catégorie associée au nom”, rather than degree modification (ibid). An investigation into the choices made by a translator whose task it was to translate the extreme degree modification inherent to the English construction will allow the relevance of these definitions to be discussed. It may be argued that according to these definitions, the English exclamative *what* will not be translated by *quel* when followed by a non-degree noun.

These aims will be achieved through a corpus study using the directional English to French sub-corpus of the Europarl corpus, a parallel corpus taken from the proceedings of the European Parliament. While the exclamative clause is has been shown to be comparatively rare (Siemund 2015), it has also been said to be a phenomenon occurring most frequently in spoken language and thus corpora of this type was deemed most appropriate for this study.

References

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

Bolinger, D. (1972). *Degree Words*. The Hague: Mouton.

Cartoni, B., Zufferey, S. & Meyer, T. (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.

Elliot, D. (1974). Towards a Grammar of Exclamations. *Foundations of Language* 11(2), 231-246.

Grevisse, M. & Goosse, A. (1990). *Nouvelle Grammaire Français*. De Boeck University.

Grevisse, M. & Goosse, A. (2008). *Le bon usage : grammaire française*. Brussels: De Boeck University.

Huddleston, R. (1984). *Introduction to the Grammar of English*. Cambridge: Cambridge University Press.

Huddleston, R. & Pullum, G. K. (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.

Michaelis, L. A. & Lambrecht K. (1996). Toward a Construction-Based Theory of Language Function: The Case of Nominal Extraposition *Language* 72(2), 215-247.

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., Greenbaum, S., Leech, G. & Svartvik, J. (1972). *A Grammar of Contemporary English*. London: Longman.

Riegel, M., Pellat, J.-C. & Rioul, R. (2009). *Grammaire méthodique du français*. Paris: PUF.

Siemund, P. (2015). Exclamative clauses in English and their relevance for theories of clause types. *Studies in Language*, 39(3), 697-727.

Siemund, P. (2017). English exclamative clauses and interrogative degree modification. In M. Napoli & M. Ravetto (eds). *Exploring Intensification: Synchronic, diachronic and cross-linguistic perspectives*. Amsterdam: Benjamins, 207-228.

Trotta, J. (2000). *Wh- Clauses in English: Aspects of theory and description*. Amsterdam: Rodopi.

Zanuttini R. & Portner P. (2003). Exclamative clauses: At the Syntax-Semantics Interface. *Language* 79(1), 39-81.

What can we learn from translation certificate examinations?
A comparison of high performing and low performing translations

Yvonne Tsai
National Taiwan University
yvtsai@ntu.edu.tw

Translation certificate examinations are designed to assess the ability of an individual to provide quality translation in a specified language combination. These exams are often administered by authorized bodies, including government organizations and translators' associations, for example, the Australian National Accreditation Authority for Translators and Interpreters, China Accreditation Test for Translators and Interpreters, the UK Institute of Translation & Interpreting, and the American Translators' Association. In 2007, the Taiwan Ministry of Education launched a Chinese and English translation and interpretation competency examination, and in 2010, it commissioned the Language Training and Testing Center to administer the exam.

The outcome of translation certificate examinations is often provided in a standardized report format, with a grade or a pass or fail result. These exams, however, can provide translation teachers and researchers with much more information than a simple letter grade. This study explored the differences between high performing translations and low performing translations by using a quantitative and qualitative analysis. By comparing the performance of candidates in translation competency exams, translation teachers can diversify and increase the efficacy of translation strategies and techniques, enabling students to apply these enhanced skills to the various situations they will encounter in their careers.

Studies have generated noteworthy findings on the features of translational corpora; however, research on corpus-based quantitative work on translation competency examinations is limited. A corpus based on English to Chinese translation certification exams is likely to generate findings similar to those of previous studies. However, would there be lexical and syntactic differences between corpora of higher and lower rated translations? If so, how can these translations enhance translation pedagogy? Through an investigation of translation competency examinations, we categorized and analyzed the common linguistic features of student translators and apply the results to improve curricula and teaching materials.

The translation exams in Taiwan consist of written translation tests featuring texts from general genres and consecutive interpretation tests. For the written translation tests, candidates can choose to take a translation exam that is either from English to Chinese (Subject A or B) or Chinese to English (Subject C or D). Subjects A and C cover various areas including business, finance, education, and culture. By contrast, Subjects B and D cover areas including science, medical care, and information technology. The texts selected for the exams are targeted to people with no expertise in the respective fields.

Written translation exams of general genre documents (Subjects A and B) were collected. After considering the number of available samples, we collected 40 random sets each of English to Chinese translations in Subject A and Subject B from 2010 to 2014. Because the exam was suspended in 2012 for development, only 320 sets were collected. The 40 sets of translations collected annually comprised 20 sets that received grades in the top 20% and another 20 that received grades in the top 40%-60%, which are referred to in this study as the "top 20% group" and the "40%-60% group," respectively. Attempts to collect translations with grades in the bottom 20% were made, but most of those were mistranslations and therefore not adopted in the analysis. On the basis of our research questions, a small learner corpus was built to analyze the translations of passive constructions and relative clauses from English into Chinese and compare the strategies adopted by referring to the overall translation quality between the top 20% group and the 40%-60% group.

The collected data were first segmented and parsed using the Chinese Word Segmentation System with Unknown Word Identification (CKIP 2017), developed by Academia Sinica, and analyzed using WordSmith

Tools 7.0 (Scott 2017). Of the three main functions of WordSmith Tools, only Concord and WordList were used. Concord presents a concordance display, collocations, patterns, and clusters of the searched word. This function places the words next to the search word and lexical patterns in the concordance. WordList shows the frequency of words and statistics from text analysis. A consistency analysis can be performed using WordList, and overlap can be identified between the vocabularies of the two wordlists.

This study examined translations submitted for translation competency exams administered by the Language Training and Testing Center. Specifically, the study analyzed the differences between translations in the high performing group and those in the low performing group to gain insights from the translation practices of the high performing group. From a pedagogical perspective, the findings of our analysis of the adequacy of strategies for translating the passive voice and relative clauses in relation to overall performance in the translation competency exam can be directly applied to student translators. By referencing the findings, student translators can learn which translation strategies should be applied in various situations, and their linguistic sensitivity can be strengthened. Therefore, the findings herein can enhance translation pedagogy and, thus, improve the translation quality of student translators.

Qualitatively, the high performing group exhibited a stronger command of translation strategies. This was represented by the diverse translation strategies they applied in their translations. The high performing group also reordered sentences in their translations, demonstrating a consideration of differences in linguistic structure. This practice was not common in the low performing group. Moreover, more mistranslations were identified in the low performing group than in the high performing group.

Considering the complexity of sentence structures that use the passive voice and relative clauses, errors in the translations of such sentences are not only common but also easily identifiable, especially in English-Chinese translations. These structures can be found in any type of text; therefore, improving the understanding of the strategies for translating the passive voice and relative clauses can reduce students’ frustration and facilitate the translation process. The passive voice and relative clauses should be emphasised in curricula. Future studies can include the low performing group in a qualitative study on the translation of both the passive voice and relative clauses for a more comprehensive view of student performance in the use of the passive voice and relative clauses.

References

CKIP. (2017, 2017/06/20). Chinese Word Segmentation System with Unknown Word Identification. Retrieved from http://ckipsvr.iis.sinica.edu.tw/.

Scott, M. (2017). WordSmith Tools (Version 7.0.0.114). Liverpool: Lexical Analysis Software. Retrieved from http://www.lexically.net/wordsmith/index.html.

Tracing the effect of pivot languages in indirect translation

Michael Ustaszewski
University of Innsbruck
michael.ustaszewski@uibk.ac.at

Background

Indirect translation, also termed relay translation (Ringmar 2013), refers to the practice of translating texts from a source language into a target language via a third language, the so-called intermediate or pivot language. While having a long-lasting tradition, especially in the case of literary translation, indirect translation has received little attention in translation studies and has only recently gained popularity as an object of study (Assis Rosa et al. 2017). In an increasingly globalised world, resorting to pivot languages has been applied to machine translation and is a widespread practice to alleviate the lack of skilled translators for less common language combinations, thus reducing costs in highly multilingual settings. The European Parliament (2008), for instance, works in 24 languages and resorts to indirect translation rather than directly translating in each of the 552 potential language combinations, with English, French and German serving as pivot languages (Katsarova 2011: 3).

The vast majority of research on indirect translation deals with literary translation, especially from a historical perspective, while research into other text types is almost non-existent (Pięta 2017: 200). What is more, most studies appear to focus on sociological and cultural aspects of indirect translation, whereas corpus-based studies aiming to elucidate the linguistic characteristics of indirect vs. direct translation are clearly underrepresented. The few existing linguistically oriented corpus studies (e.g. Zubillaga Gomez 2016) are rather qualitative in nature and employ basic univariate statistics. In light of the (largely economically motivated) importance of indirect translation in the language industry, addressing the identified research gap is a desideratum with relevance for translation theory and practice alike.

Aims & research question

The aims of this contribution are twofold. Firstly, and in line with quantitative research on variation in translation (e.g. Diwersy et al. 2014; Lapshinova-Koltunski 2017; Lapshinova-Koltunski & Zampieri 2017), we aim to explore weather hierarchical cluster analysis is able to detect differences between direct translations vs. indirect translations vs. non-translated originals. Secondly, exploratory multivariate techniques are to be employed in order to investigate the influence of pivot languages on translations. The underlying research question thus reads as follows: “Is there an effect of the pivot language on target texts in indirect translation, and is this effect strong enough to discriminate between direct and indirect translations?” The overarching goal is to provide empirically grounded insights into this special and underresearched yet common form of mediated communication. At the same time we extend the scope of research into translationese from two languages interacting in the translation process to three languages.

Methodology

For our analyses we automatically extracted all comparable corpora implicitly contained in the Europarl corpus (Koehn 2005) using the *EuroparlExtract* toolkit (Ustaszewski 2018). Europarl includes translations of the debates at the European Parliament in 21 languages from 1996 to 2012. According to Cartoni and Meyer (2012: 2134), translations produced after 2003 are indirect, whereas before 2003 translations were made directly without pivot languages, at least for the more widely used languages. The availability of both direct and indirect translations within one single corpus that is linguistically diverse yet topically homogenous makes Europarl a prime candidate to investigate pivot language effects in translation. Following common practice in research into translationese and the characteristics of translated language, our study is based on monolingual comparable corpora. However, the comparable corpora we extracted from Europarl contain not only two (translated vs. non-translated) but three (directly translated vs. indirectly translated vs. non-translated) text production conditions for each of the 21 languages. For data analysis, all texts in our corpus are represented as multidimensional vectors of linguistically motivated content-independent features (function word frequencies, lexical density and lexical richness, readability scores, part-of-speech n-grams, mean word ranks, repetitions, etc.). Subsequently, hierarchical cluster analysis is to be performed in order to explore (dis-)similarities among texts. The reason why an unsupervised machine learning approach is given preference over supervised ones is that the Europarl corpus does not contain explicit metadata about the direct vs. indirect status of translations and therefore it is not entirely sure which texts belong to which category (Cartoni & Meyer 2012: 2134). In order to gain more meaningful insights into the complex interplay of the observed features and their role in potential pivot language effects, Principal Component Analysis will be conducted. Both analyses are to be computed with the software package *Stylo* (Eder et al. 2016), which is implemented in R and widely used in computational stylometry. If the exploratory analyses reveal strong pivot language effects, a validation of the findings will be carried out by means of supervised machine-learning classifiers, which is in line with previous research on the automatic detection of translationese (e.g. Baroni & Bernardini 2006; Volansky et al. 2015). This complementary step is to be carried out using the Weka machine learning toolkit (Frank et al. 2016).

Preliminary results and outlook

The comparable corpora we extracted from Europarl with *EuroparlExtract* comprise over 500 million tokens across all 21 languages in the translated and over 40 million tokens in the non-translated section, which is deemed to be a sufficiently large data set to investigate pivot language effects in various language families. To the best of our knowledge, our work is going to be the first multivariate corpus-based analysis of indirect translation in the non-literary genre. This is expected not only to strengthen the understanding of the intricate nature of translationese, but also to pave the way for more rigorous research in the relatively young and unexplored area of research into indirect translation. As a matter of fact, the main limitation of our research design, i.e. the incompleteness of explicit information about the factual translational status of texts in our corpus, adds an interesting investigative perspective to our study: sophisticated data analysis techniques may help to clarify under what circumstances particular portions of the Europarl corpus have been produced and thus to complement important but missing meta-information about the composition of this influential linguistic resource.

References

Assis Rosa, A., Pięta, H. & Bueno Maia, R. (2017). Theoretical, methodological and terminological issues regarding indirect translation: An overview. *Translation Studies* 10(2), 113-132.

Baroni, M. & Bernardini, S. (2006). A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing* 21(3), 259-274.

Cartoni, B. & Meyer, T. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2132-2137.

Diwersy, S., Evert, S. & Neumann, S. (2014). A Weakly Supervised Multivariate Approach to the Study of Language Variation. In B. Smrzecsanyi & B. Wälchli (eds). *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. Berlin: De Gruyter, 174-204.

Eder, M., Rybicki, J. & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal* (8)1, 107-121. https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf.

European Parliament (2008). *European Parliament – never lost in translation*. Press relase by the Directorate for the Media. http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+IM-PRESS+20071017FCS11816+0+DOC+PDF+V0//EN&language=EN.

Frank, E., Hall, M. A. & Witten, I. H. (2016). *The WEKA Workbench*. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques'.

Katsarova, I. (2011). *The EU and multilingualism*. http://www.europarl.europa.eu/RegData/bibliotheque/briefing/2011/110248/LDM_BRI(2011)110248_REV1_EN.pdf

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit,* Phuket, Thailand, 79-86.

Lapshinova-Koltunski, E. (2017). Exploratory Analysis of Dimensions Influencing Variation in Translation: The case of text register and translation method. In G. De Sutter, I. Delaere & M.-A. Lefer (eds). *Empirical Translation Studies. New Theoretical and Methodological Traditions*. Berlin: Mouton de Gruyter, 207-234.

Lapshinova-Koltunski, E. & Zampieri, M. (2017). Linguistic Features of Genre and Method Variation in Translation: A Computational Perspective. In D. Legallois, T. Charnois & M. Larjavaara (eds). *Grammar of Genres and Styles*, Berlin: De Gruyter, 92-117.

Pięta, H. (2017). Theoretical, methodological and terminological issues in researching indirect translation: A critical annotated bibliography. *Translation Studies* 10(2), 198-216.

Ringmar, M. (2012). Relay translation. In Y. Gambier & L. van Doorslaer (eds). *Handbook of Translation Studies* (Vol. 3). Amsterdam: John Benjamins, 141-144.

Ustaszewski, M. (2018). *EuroparlExtract*. Open-source software. https://github.com/mustaszewski/europarl-extract.

Volansky, V., Ordan, N. & Wintner, S. (2015). On the features of translationese. *Literary and Linguistic Computing* 30(1), 98-118.

Zampieri, M. & Lapshinova-Koltunski, E. (2015). Investigating Genre and Method Variation in Translation Using Text Classification. In P. Král & V. Matoušek (eds). *Proceedings of the 18th International Conference on Text, Speech & Dialogue*. Cham: Springer, 41-50.

Zubillaga Gomez, N. (2016). (In)direct offense. A comparison of direct and indirect translations of German offensive language into Basque. *Perspectives* 24(3), 486-497.

Lexical and morphological features of translational Lithuanian

Jurgita Vaičenonienė, Jolanta Kovalevskaitė
Vytautas Magnus University
jurgita.vaicenoniene@vdu.lt, jolanta.kovalevskaite@vdu.lt

The primary aim of this presentation is to investigate the lexical and morphological features of Lithuanian translated from English using the methods of corpus linguistics. For this purpose, a Comparable Corpus of Original and Translated Lithuanian (ORVELIT) was used. The corpus consists of four subcorpora of original and translated fiction and popular science literature (1 million words each) (Vaičenonienė et al. 2017). Lithuanian, in which grammatical functions are expressed by endings and inflectional suffixes rather than function words or word order, is characterized by a high morphological ambiguity. For example, Rimkutė (2006) identified that 47% of word forms were morphologically ambiguous in a 1-million-word corpus of general Lithuanian. Therefore, raw corpus may not always be helpful to retrieve more precise information about the language. ORVELIT was automatically morphologically annotated with *Semantika.lt* analyser.[1] Previous research has shown that this tool achieves a precision of ~98.0%, ~95.3%, and ~86.8% of accuracy on the lemmatization, POS tagging and the annotation of the morphological categories, respectively (Kapočiūtė-Dzikienė et al. 2017). The following research objectives have been set:

1) to compare original and translated Lithuanian in the raw corpus version focusing on type token ratios, sentence length and high frequency words;
2) to discuss the distribution of parts of speech and morphological categories in original and translated fiction and popular science literature in the morphologically annotated version of the corpus.

The general statistics function of the WordSmith Tools 7.0 (Scott 2016) has allowed to compare the standardized type token ratios and mean sentence length in originals and translations. Original Lithuanian fiction (66.02%) appears to be lexically richer in comparison to translated (63.14%), whereas the standardized type token ratios of original and translated popular science are rather similar (62.38% vs. 63.53%). Previous research has also shown that translations do not always have a lower type token ratio and that this parameter may depend on the language pair or register in question (e.g. Xiao & Dai 2014). Differently from the findings of Laviosa's (1998) research, mean sentence length in words tends to be higher in original fiction (10,85 vs. 8,75) and popular science (15,55 vs. 13,36) texts, which may indicate the tendency to simplify or split syntactically complex sentences in translation. Another commonly tested parameter to show that translations are prone to lexical simplification is lexical density. For example, Laviosa (1998) has found that in English, high frequency words or "words with a minimum percentage of 0.1% of the total corpus" (Xiao & Dai 2014: 7) (1) constitute a larger part of the word list than low frequency words and (2) their proportion is higher in translations. As the results show, because of the inflectional nature of Lithuanian, list heads do not account for a greater half of the corpus as in English. However, in agreement with Laviosa's (1998) and Xiao & Dai's (2014) research, the proportion of high frequency words in comparison to low frequency words is higher in translations (30.53% vs. 33.35% for original and translated fiction; 24.26% vs. 29.35% for original and translated popular science).

Analysis of the distribution of parts of speech in the morphologically annotated version of the corpus shows that content words (nouns, verbs, adjectives and adverbs) make up a higher proportion of the word list in the four subcorpora (63% vs. 61% for original and translated fiction and 68% vs. 66% for original and translated popular science). Also, the proportion of the content words is higher in original Lithuanian and this difference is statistically significant according to the log likelihood test (fiction: LL = 187.54 for 1 d.f.; popular science: LL = 258.27 for 1 d.f.). Nouns and adjectives are more frequent in original Lithuanian, whereas there are more verbs in translated fiction and popular science, which, at first sight, does not support the often-cited claim that the verbal nature of Lithuanian tends to be underrepresented when translating from English. On the other hand,

[1] http://semantika.lt/SyntaticAndSemanticAnalysis/Analysis.

a more detailed analysis of verb patterns and their syntactic features as, for example, complex predicates, is necessary to investigate the verbal specificities of translated Lithuanian. Verb pattern differences in translational Lithuanian are also evidenced by lower participle and higher infinitive frequencies in comparison to original Lithuanian in both registers. Adverbs have lower frequencies in translated fiction, but higher in translated popular science. Although the major classes of function words (pronouns, numerals, particles, conjunctions, interjections and prepositions) occur with a statistically significant higher frequency in translations (fiction: LL = 442.71 for 1 d.f.; popular science: LL = 3672.20 for 1 d.f.), there is a variation in their spread. Especially outstanding is a statistically significant overuse of the largest category of function words, pronouns, in translations (fiction: LL = 1082.34 for 1 d.f.; popular science: LL = 6912.32 for 1 d.f.).

In sum, initial results show that translations from English deviate in certain ways from original Lithuanian. Further research will provide a more detailed discussion of the distribution of parts of speech, their morphological categories (cases, verb, pronoun and adjective types and tenses) as well as register determined frequency differences.

References

Corpus of the Contemporary Lithuanian Language. http://tekstynas.vdu.lt/tekstynas/ (Accessed 01 10 2018).

Krause, T. & Zeldes, A. (2016). ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities* (31). http://dsh.oxfordjournals.org/content/31/1/118 (Accessed 01 10 2018).

Kapočiūtė-Dzikienė, J., Rimkutė, E. & Boizou, L. (2017). A Comparison of Lithuanian Morphological Analyzers. In K. Ekštein & V. Matoušek (eds). *The Proceedings of the TSD 2017: Text, Speech, and Dialogue: 20th International Conference*. Springer, 47-56.

Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta,* 43(4), 557-570.

Parallel Corpus. http://tekstynas.vdu.lt/page.xhtml?id=parallelCorpus (Accessed 01 10 2018).

Rimkutė, E. (2006). *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne [Morphological Disambiguation in Computerized Corpora].* PhD Thesis. Kaunas: Vytautas Magnus University.

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

Vaičenonienė, J., Kovalevskaitė, J. & Ringailienė, T. (2017). Tekstynais paremti vertimų kalbos tyrimai ir šaltiniai. *Kalbų studijos* 30, 42-55.

Xiao, R. & 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.

Questioning explicitation in translation studies: A multifactorial corpus investigation of the *om*-alternation in translated and original Dutch

Amélie Van Beveren, Gert De Sutter, Timothy Colleman
Ghent University
Amelie.VanBeveren@ugent.be, Gert.DeSutter@ugent.be, Timothy.Colleman@ugent.be

Since Blum-Kulka's (2000 [1986]) Explicitation Hypothesis, explicitation is known as one of the many features associated with translated language. Although most studies agree that explicitation occurs more often in translations than in non-translated texts (see e.g. Øverås 1998; Olohan & Baker 2000), the causes of the increased explicitness in translations have not received a lot of empirical attention. In a recent study, Kruger & De Sutter (to appear) try to disentangle the three most common explanations proposed for the explicitation of complementizer *that* in English through a multifactorial corpus investigation: the processing-strain hypothesis (Olohan & Baker 2000), the risk-aversion hypothesis (Pym 2005, 2015) and the source-language transfer hypothesis (Becher 2010). Kruger & De Sutter (to appear) rule out cross-linguistic interference and emphasize the importance of processing effort and risk avoidance as the mechanisms behind explicitation in translation. Register has repeatedly been shown to influence the likelihood of complementizer omission (De Sutter et al. 2012). Kruger (2018) proved that translations demonstrate less sensitivity to register, because choosing more explicit encodings in translation is a solution that reduces risk independent of the register of the text. As regards processing effort, Kruger & De Sutter (to appear) hypothesize that "the demands of processing resources may lead translators to select the default option as a way of reducing effort" (Kruger & De Sutter to appear: 27). In other words, they conclude that translations are seen as cognitively more demanding environments and a higher processing cost is associated with the choice of the default or the highest frequent option, which in their case study is the explicit form. But what happens with cases in which the default option is not the explicit form, but the implicit one? Do we still find more explicit forms in translated texts or does the translator indeed choose for the default – in this case implicit – option?

In this case study, we address those questions by zooming in on a particular case of grammatical alternation, viz. the variation between infinitival complements (=IC) with and without the prepositional complementizer *om* in Dutch, illustrated in (1) below, where the infinitival clause depends on a verb, a noun, and an adjective, respectively. When *om* is present, it functions as an explicit boundary signal and it clarifies the clause structure. Descriptive and prescriptive grammars state that the implicit form is more common in written language and the explicit form in spoken language (ANS 1997, Syntax of Dutch 2015). In other words, the implicit form is assumed to be the default option for both writers and translators. Some examples of the alternation:

(1) a. *Hij beslist (om) thuis te blijven* 'He decides to stay at home'
b. *Zijn belofte (om) op tijd te komen* 'His promise to be on time'
c. *Ik ben bang (om) je kwijt te raken* 'I am afraid to lose you'

Previous empirical analyses with non-translated data have pointed out that the complementizer *om* can be added or omitted depending on different syntactic, semantic and pragmatic factors (ANS 1997; Vliegen 2001; SOD 2015). Many translation scholars nowadays realize that linguistic behavior in translations should be analyzed as a multifactorial phenomenon rather than a monofactorial one (De Sutter & Lefer 2016; Kruger 2018). Like Kruger & De Sutter (to appear), we test the three proposed explanations for increased explicitness by taking into account variables as source language for the source-language transfer hypothesis, register for the risk-aversion hypothesis and complexity-related factors for the processing-strain hypothesis. Some examples of factors dealing with complexity are the distance between the matrix-clause verb and the onset of the IC, the syntactic complexity of the IC, the modality of the verb (active vs. passive, finite vs. infinite) and the syntactic level of the IC (coordination/subordination).

We investigated the possible effect of the above-mentioned factors on the *om*-alternation through a mixed-effects logistic regression analysis applied to a database of real-language examples culled from the Dutch

Parallel Corpus, a multiregister and bidirectional parallel corpus of Dutch, English and French with Dutch as a central language. On the basis of three different regression analyses (one for both translated and non-translated texts and one for translated texts and non-translated texts separately), we found that translations have a clear preference for the implicit form. This finding refutes the idea that translated texts systematically exhibit a preference for explicit forms, and confirms Kruger & De Sutter's hypothesis that translators might have a preference for the most frequently used forms. The other results are in line with the conclusions of both Kruger (2018) and Kruger & De Sutter (to appear): source language interference does not play a significant role, but the effect of complexity-related factors does, both in translated and non-translated texts. A final interesting result is that register only plays a significant role in the model based on non-translated texts. Translations do not take into account the specific preference of a register for explicit or implicit *om*. For that reason, we consider the risk-avoidance hypothesis as the most convincing mechanism for explaining the different behavior of translators while translating texts except that avoiding risk is not a priori the same as choosing the explicit option. Avoiding risk is also possible by opting for the most frequent form of the construction, even if that happens to be the more implicit form.

References

ANS = Haeseryn, W. et al. (1997). *Algemene Nederlandse Spraakkunst*. Groningen: Nijhoff, Deurne: Wolters Plantyn.

Becher, V. (2010). Abandoning the notion of "translation-inherent" explicitation. Against a dogma of translation studies. *Across Languages and Cultures* 11(1), 1-28.

Blum-Kulka, S. (2000[1986]). Shifts of Cohesion and Coherence in Translation. In L. Venuti *The Translation Studies Reader* (1st ed.). London: Routledge, 298-313.

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

De Sutter, G., & Lefer, M.-A. (2016). Empirical Translation Studies in the Post-Baker Era: Towards a New Research Agenda. Unpublished conference paper presented at the 8th EST Congress, 15-17 September, Aarhus.

Kruger, H. (2018). *That* Again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English. *Across Languages and Cultures*.

Kruger, H. & De Sutter, G. (to appear). Alternations in contact and non-contact varieties: Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach.

Olohan, M. & Baker, M. (2000). Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1(2), 141-158.

Øverås, L. (1998). In search of the third code: An investigation of norms in literary translation. *Meta* 43(4).

Pym, A. (2005). Explaining explicitation. In K. Karoly & A. Foris (eds). *New Trends in Translation studies: In Honour of Kinga Klaudy*. Budapest: Akademiai Kiado, 29-34.

Pym, A. (2015). Translating as Risk Management. *Journal of Pragmatics* 85, 67-80.

SOD = Broekhuis, H., Corver, N., Vos, R. & Bennis, H. (2015). *Syntax of Dutch: verbs and verb phrases*. Amsterdam: Amsterdam University Press.

Vliegen, M. (2001). Het facultatieve *om* na illocutionaire werkwoorden. *Nederlandse taalkunde* 6, 112-132.

Lexical knowledge and translation self-revision behaviour as revealed in a trainee corpus, eyetracking and keystroke logging data

Sonia Vandepitte
Ghent University
sonia.vandepitte@ugent.be

Since the 1990s translation training methods have seen a wide range of pedagogical developments that are oriented towards learner autonomy, such as Kiraly's social constructivist and his later emerging theory (Kiraly 2000, 2003) or Flanagan and Heine's scaffolded approach to peer feedback (Flanagan & Heine 2015). Nevertheless, research into what has been called 'self-directed learning routines' (Albin 2014: 98) has remained scarce.

In translation, learner autonomy can be understood to be revealed in the final translations of a trainee - the more learner autonomy is demonstrated, the better translations are expected - as is implied in many studies. However, the learning process itself, i.e. the change from a stage in which an infelicity is produced to the stage in which the learner is aware that a text contains an infelicity and that he/she can even amend the passage, sometimes becomes visible itself if the self-revision behaviour of the trainee is investigated by means of process-oriented research methods such as eyetracking and keystroke logging.

Self-revision in the translation context has already been studied empirically by e.g. Mossop (2014), Robert (2008, 2012), Robert et al. (2016). Robert (2012), for instance, found that the process of self-revision is influenced by the procedure adopted, whether the latter is monolingual, bilingual, bilingual followed by monolingual, or monolingual followed by bilingual revision. Schaeffer, Tardel, Hofmann, and Hansen-Schirra (in press), observe an effect of revision behaviour on revision duration: the most efficient process is the one in which much reading and writing take place simultaneously applying few deletions during the drafting phase.

Both studies focus on the effect that revision behaviour has on the translation product/process, implying that self-revision behaviour needs to be understood as an important factor in the production of translations. Finding out which factors play a role in self-revision behaviour therefore also looks like a promising avenue to inform not only the researchers about the translation process but also translation trainers that engage in learner autonomy and organize self-revision activities in their classes.

Hence, the present research will show the results of an inquiry into the relationship, if any, between one aspect of the trainees' linguistic competence and revision behaviour. In particular, it will look into whether there is a relation between the trainees' level of lexical knowledge as laid down in the LexTALE test (Lemhöfer & Broersma 2012) and their revision behaviour, the latter being defined as the frequency with which a trainee identifies and changes a word, phrase or passage into a more adequate and/or acceptable one:

- Is there a relation between better lexical knowledge and better self-revision behaviour?

For translation trainers, it would also be good to know whether the students with better lexical knowledge not only focus on lexical issues but also on grammatical, textual and translational ones, which leads to the following two questions:

- Is there a relation between better lexical knowledge and the content which self-revision focuses upon (in terms of lexical versus all other textual issues)?
- Is there a relation between better lexical knowledge and the different revision types as identified in Angelone (in press), to wit deletions, additions and changes?

The data for this presentation has been collected from an experimental setting of translation sessions which includes Translog recordings of sixteen participants who each participated in five short sessions, translating a number of short passages from English into Dutch (L2 into L1). While the presentation will discuss how the two different types of data can be triangulated with each other – how do the learner-translated text corpus data (their translations) inform keystroke logging behavioural data and vice versa, which information is contributed by the process data to the corpus data – the replies to the research questions above will also reveal any differences in translation self-revision between beginning translation trainees (first-year bachelor students) and the more advanced ones (master students).

References

Albin, J. (2014). *The Reflective Translator: Strategies and Affects of Self-Directed Professionals.* Frankfurt am Main: Peter Lang.

Angelone, E. (in press). In E. Huertas Barros, S. Vandepitte & E. Iglesias Fernández (eds). *Quality Assurance and Assessment Practices in Translation and Interpreting*. Hershey, PA: IGI Global.

Flanagan, M. & Heine, C. (2015). Peer-feedback as a translation training tool in web-based communication. *Hermes* 54, 115-136.

Kiraly, D. (2000). *A social constructivist approach to translator education: Empowerment from theory to practice*. Manchester: St Jerome Publishing.

Kiraly, D. (2003). From Teacher-Centred to Learning-Centred Classrooms in Translator Education: Control, Chaos or Collaboration? In A. Pym, C. Fallada, J. R. Biau & J. Orenstein (eds). *Innovation and E-Learning in Translator Training: Reports on Online Symposia*. Tarragona: Universitat Rovira I Virgili, 27-31.

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44, 325-343.

Robert, I. S. (2008). Translation revision procedures: An explorative study. Retrieved from http://www.arts.kuleuven.be/cetra/papers/files/robert.pdf.

Robert, I. S. (2012). *La révision en traduction : les procédures de révision et leur impact sur le produit et le processus de révision* (Unpublished PhD thesis). University of Antwerp, Antwerpen.

Schaeffer, M., Tardel, A., Hofmann, S. & Hansen-Schirra, S. (in press). Cognitive effort and efficiency in translation revision. In E. Huertas Barros, S. Vandepitte & E. Iglesias Fernández (eds). *Quality Assurance and Assessment Practices in Translation and Interpreting*. Hershey, PA: IGI Global.

Keen on cognates or afraid of false friends?
Cognate ratios in translated and non-translated Dutch

Lore Vandevoorde, Els Lefever
Ghent University
lore.vandevoorde@ugent.be, els.lefever@ugent.be

Introduction

The use of cognate translations is an intriguing topic in Translation Studies (TS), but under-researched from the point of view of Corpus-based TS. Most research in the area of cognate translations either concerns interpreting studies (Dam 1998; Christoffels 2004; Defrancq 2015) and/or is process-oriented in nature, i.e. focusing on the circumstances which will prompt translators to choose a cognate translation over a non-cognate translation (Shlesinger & Malkiel 2005; Malkiel 2009; Oster 2017). From a product-based perspective, Hansen-Schirra et al. (2017) investigated *inter alia* the influence of translation-inherent constraints such as context and language status and found that some languages are more receptive to the use of cognates than others, and that different text types have different preferences with respect to cognate use. From a purely corpus-based perspective, however, not much is known about the actual proportion of cognate words in translated texts compared to the proportion of cognate words in originally written texts. To our knowledge, only two studies have addressed this matter in a corpus-based fashion. First, Gellerstam (1986) concluded that translated texts (using a corpus of Swedish novels translated from English) contain more cognates than non-translated texts. Vintar & Hansen-Schirra (2005) found that German is more receptive for cognates than Slovene, both in original texts and in translated texts. With the current study, we aim to provide some additional insights into the use of cognate translations by answering two basic, yet unresolved corpus-based questions: (i) do translated texts (translated into Dutch) contain more cognate words than non-translated texts (text originally written in Dutch)? and (ii) is the cognate ratio in translated texts from more cognate languages (e.g. English and Dutch) higher than the cognate ratio in translated texts from less cognate languages (e.g. French and Dutch)?

Hypotheses

Depending on the hypothesis towards which one is more favorably disposed, different outcomes can be expected. First, following Halverson's (2015: 320) idea of default translation – and under the assumption that the cognate translation is the default, meaning that it is unconstrained and immediately produced – we can expect translated texts to contain more cognate words than non-translated texts. Second, and corresponding to the cognate facilitation effect hypothesis, we expect texts translated between more cognate languages (e.g. source English and target Dutch) to contain more cognate words than texts translated between less cognate languages (e.g. source French and target Dutch). From psycholinguistic experiments, there is overwhelming evidence for a cognate facilitation effect (bilinguals are faster and more accurate at producing cognates than control words which only exist in one of their languages) (Schepens et al. 2012, 157-158). This effect then suggests that translators, when confronted with an L2 source language word which has a cognate translation in the L1 target language, will be more likely to choose the cognate translation over a non-cognate equivalent. Since higher cognateness between two languages will create more occasions in which the cognate facilitation effect could facilitate the production of a cognate translation, the expectation on the basis of the cognate facilitation effect would be that the higher the level of cognateness between a source and target language, the higher the ratio of cognates in translated texts will be. Thirdly, and in accordance with the fear of false friends hypothesis, we can expect exactly the opposite outcome: translated texts from more cognate languages (e.g. English and Dutch) will contain fewer cognate words than translated texts from less cognate languages (e.g. French and Dutch). The fear of false friends hypothesis is based on the results of Kussmaul (1995), Kussmaul & Tirkkonen-Condit (1995) and more recently Shlesinger & Malkiel (2005) and Malkiel (2009) who concluded that, out of a hypothesized "fear of false friends", translators tend to choose a non-cognate translation over a cognate translation when both are (presumably) equally translationally equivalent. In other words, the third hypothesis leads to the expectation that the higher the level of cognateness between the source and the target language, the lower the ratio of cognate words in translated texts since high cognateness between languages will increase the translators' fear of false friends and lead to more non-cognate translations.

Method

In order to test the three hypotheses, we compared the English-Dutch and French-Dutch cognate ratios for different sub-corpora of the Dutch Parallel Corpus, a ten-million word, sentence aligned parallel corpus which contains balanced and comparable sub-corpora of originally written Dutch (OrDutch; token size = 4,911,944), Dutch translated from French (TDFrench; token size = 2,076,443) and Dutch translated from English (TDEnglish; token size = 2,539,248) (Macken et al. 2011). The token frequencies of the Dutch word types from a list of Dutch-French cognate words established by Schepens et al. (2013) (a more exhaustive, self-compiled list of Dutch-French cognate words is currently under construction) (n= 559) were looked up in each of the sub-corpora (OrDutch, TDFrench and TDEnglish) and divided by the number of tokens in each sub-corpus. The same was done for the token frequencies of the Dutch word types from a list of Dutch-English cognate words, equally established by Schepens and colleagues (n=1104). The ratios of French-Dutch cognate word tokens (n = 78,669, cognate word token ratio = 0.0160) and English-Dutch cognate word tokens (n= 226,900, cognate word token ratio = 0.0462) in non-translated Dutch texts (sub-corpus = OrDutch) were taken as respective points of comparison for TDFrench and TDEnglish.

Results

The resulting cognate ratios ($\frac{\text{cognate}-\text{token frequency}}{\text{total token frequency}}$) show that the ratio of English-Dutch cognate words in TDEnglish (cognate word token ratio = 0,0416) is significantly *lower* than in OrDutch (χ^2= 763.46, df = 1, p= < 0.001). This means that there are significantly fewer Dutch-English cognate words in Dutch texts translated from English than in texts originally written in Dutch. For TDFrench, we find that the ratio of Dutch-French cognate words is significantly *higher* in TDFrench (cognate word token ratio = 0.0249) than in OrDutch (χ^2= 6,023.9, df = 1, p < 0.001), meaning that in TDFrench, there are significantly more Dutch-French cognate words compared to OrDutch. These results argue in favor of the fear of false friends hypothesis: more cognate languages (English-Dutch) exhibit lower cognate ratios in translated language (translated Dutch) (out of a "fear of false friends") than less cognate languages (French-Dutch) where, due to low cognateness between the languages, fear of false friends does not operate. In order to confirm the fear of false friends hypothesis, more evidence is needed from other language pairs (e.g. English texts translated from Spanish vs. English texts translated from German compared to texts originally written in English, or language pairs including non-Indo-European languages), especially since previous research by Hansen-Schirra and colleagues has shown that cognate-receptiveness is language-specific.

References

Christoffels, I. K. (2004). *Cognitive Studies in Simultaneous Interpreting*. PhD thesis, University of Amsterdam.

Defrancq, B. (2015). Corpus-Based Research into the Presumed Effects of Short EVS. *Interpreting* 17(1), 26-45.

Gellerstam, M. (1986). Translationese in Swedish Novels Translated from English. In L. Wollin & H. Lindquist (eds). *Translation Studies in Scandinavia. Poceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*. Lund Studies in English. Lund: CWK Gleerup, 88-95.

Halverson, S. (2015). Cognitive Translation Studies and the Merging of Empirical Paradigms. The Case of 'literal translation'. *Translation Spaces* 4(2), 310-340. doi:10.1075/ts.4.2.07hal.

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*. Translation and Multilingual Natural Language Processing 7. Berlin: Language Science Press, 3-22.

Kussmaul, P. (1995). *Training the Translator*. Amsterdam & Philadelphia: John Benjamins.

Kussmaul, P. & Tirkkonen-Condit, S. (1995). Think-Aloud Protocol Analysis in Translation Studies. *TTR* 8(1), 177-199.

Malkiel, B. (2009). Translation as a Decision Process. Evidence from Cognates. *Babel* 55(3), 228-243.

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*. Translation and Multilingual Natural Language Processing 7. Berlin: Language Science Press, 23-39.

Schepens, J., Dijkstra, T. & Grootjen, F. (2012). Distributions of Cognates in Europe as Based on Levenshtein Distance. *Bilingualism: Language and Cognition* 15(1), 157-166. doi:10.1017/S1366728910000623.

Schepens, J., Dijkstra, T., Grootjen, F. & van Heuven, W. J. B. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLoS ONE* 8(5), e63006. doi:10.1371/journal.pone.0063006.

Shlesinger, M. & Malkiel, B. (2005). Comparing Modalities: Cognates as a Case in Point. *Across Languages and Cultures* 6(2), 173-193.

Dam, H. V. (1998). Lexical Similarity vs Lexical Dissimilarity in Consecutive Interpreting. A Product-Oriented Study of Form-Based vs. Meaning-Based Interpreting. In F. Pöchhacker & M. Shlesinger (eds). *The Interpreting Studies Reader*. London & New York: Routledge, 267-276.

Vintar, S. & Hansen-Schirra, S. (2005). Cognates: Free Rides, False Friends or Stylistic Devices? A Corpus-Based Comparative Study. In G. Barnbrook, P. Danielsson & M. Mahlberg (eds). *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. London & New York: Continuum, 208-221.

Is target text quality indicative of translation and translator competence? A bi-national study on learning progress in a diversified learner translator corpus

Heidi Verplaetse[1], Gys-Walt van Egdom[2], Iris Schrijver[3],
Winibert Segers[1], Hendrik Kockaert[1], Fedde van Santen[4]
KU Leuven[1], Zuyd University of Applied Sciences[2], University of Antwerp[3],
ITV University of Applied Sciences for Translation and Interpreting[4]
heidi.verplaetse@kuleuven.be, gijs-walt.vanegdom@zuyd.nl, iris.schrijver@uantwerpen.be,
winibert.segers@kuleuven.be, hendrik.kockaert@kuleuven.be, feddevansanten@gmail.com

At a time when competence research was rapidly gaining momentum in Translation Studies, Adab emphasised that '[i]n the context of developing translation competence, one of the questions to be considered is that of how to evaluate the target text, *as product of the process*' (2000: 215, emphasis added).' It seems safe to say that but few scholars would have been inclined to challenge this claim at that time. Even today, the claim would be considered little more than a trite observation. In translator training, the level of competence of trainee translators is often gauged and assessed through translation exercises in which the trainee is instructed to produce a target text. Trainees are believed to have reached a certain level of competence/to have acquired a competence, if the quality of the abovementioned text comes up to standard.

We argue that, by accepting and adopting this view, the view that translation quality is the ultimate basis on which to ground an evaluative judgment on the competence levels of (aspiring) translators, one inevitably sidesteps a few crucial and incontestably thorny issues. First of all, one should note that, from a purely professional point of view, translation is considered both a collaborative activity (in which several language professionals, and sometimes even domain specialists and clients, work toward a common goal) and a service (ISO 17100, 2015; cf. EMT Expert Group 2009, 2017). By taking the translation of an individual student as a point of departure in the assessment of translation competence, one overlooks the collaborative and service-related aspects of translation. The second argument against an uncritical equation of translation quality and translation/ translator competence, relies on the idea that, even if an all too narrow definition of translation competence were to be applied, subcompetences ought be tested in a reliable and a valid manner. This means that, in order for translation competence to be measured, translator trainers should be able to identify text items in the target text, items that attest to translation behaviour that is in line with very specific "can-do" statements that have been proven (beyond doubt) to be indicative of very specific subcompetences, without there being the possibility that the same item attests to behaviour that is directly or indirectly associated with other "can-do" statements or other subcompetences. With translation competence models geared to a translation market in constant flux, prospects do not look auspicious for measurement of translation/translator competence on the meager basis of textual material.

By far the most important objection that can be made against the assessment of trainee translator competence through translation product evaluation, is the plain fact that the correlation between the two has never been scientifically tested/corroborated, in spite of the fact that the gut-feeling tells us that translation quality reflects translation/translator competence. The research project presented in this paper aims to be a first audacious step in the direction of establishing the strength (or weakness) of the correlation.

In 2018, trainee translators at different institutions (Zuyd University of Applied Sciences / ITV University of Applied Sciences, KU Leuven and University of Antwerp) were asked to produce a Dutch translation of an English, a French or a Spanish source text, with Inputlog (keystroke logging software) running in the background. Different groups of students with different levels of translation experience, from first year bachelor students until fourth year master students, translated texts containing around 340 words on the topic of health economics. This yielded a student translator corpus with mixed diversification. The current corpus consists of approximately 50,000 words translated by 148 students from bachelor courses in Translation and in Applied Linguistics, and Master Translation at the abovementioned institutions. Currently, the subcorpora per training level represent respectively 16.2%, 10.1%, 54.1% and 19.6% of the full corpus for 1st, 2nd, 3rd and 4th year students. The third bachelor year represents a pivotal time for many students in the above mentioned programmes. In this manner this subcorpus may serve as a developmental benchmark. Students from the different study years translating from the same source language translated the same source text. All students' translation activity was logged in idfx files (using Inputlog), which provide data on students' translation behaviour (cf. below).

Upon completion of the assignment, the target texts were uploaded to the translation revision and evaluation platform translationQ. In the first stage of the project, the quality of the translation was evaluated. The method of choice was the preselected items evaluation (or PIE) method, a method that ensures a speedy and highly reliable evaluation of translation quality — especially when the method is employed within the translationQ environment (cf. Kockaert & Segers 2014, 2017; Van Egdom et al. forthcoming). Once the evaluation was completed, a selection was made of translations pertaining to the top and the bottom groups in terms of quality. In the second stage of the project, the processes underlying the selected translations were closely examined with a view to distinguishing translation styles (behavioural patterns) that testify to a trainee translator's competence or (relative) incompetence. Particular attention was paid to differences in revision behaviour (number of deletions, substitutions, additions, position in the target text and time in process), use of external sources (number, time spent, time in process, distribution over process) and production fluency (ratio of number of characters produced relative to total process time and number of characters in the final product), and pausing behaviour (number, duration, location, distribution over process).

In this paper, the focus will be on learning progress as attested for a diversified learner translator corpus by the keystroke logging software. As the data (viz. target texts and process data) gathered for this project were produced by students in various stages of training, one of the aims of this study was to find out whether the learning progress could be observed in the (quality of the) translation text and in the translation behaviour of trainees.

References

Beeby, A. (2000). Evaluating the Development of Translation Competence. In C. Schäffner & B. Adab (eds). *Developing translation competence.* Amsterdam: John Benjamins, 185-198.

EMT Expert Group (2009). Competences for professional translators, experts in multilingual and multimedia communication. Available: https://ec.europa.eu/info/resources-partners/european-masters-translation-emt/european-masters-translation-emt-explained_en.

EMT Expert Group (2017). *EMT competence framework.* Manuscript.

ISO-17100 (2015). *ISO 17100 Translation services: Requirements for translation services*. Geneva: International Organization for Standardization.

Kockaert, H. & Segers, W. (2014). Évaluation de la traduction : La méthode PIE (Preselected Items Evaluation). *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.

Van Egdom, G., Verplaetse, H., Schrijver, I, Kockaert, H., Segers, W., Pauwels, J., Wylin, B. & Bloemen, H. (forthcoming). How to put the Translation Test to the Test? On Preselected Items Evaluation and Perturbation. In E. Huertas Barros, S. Vandepitte & E. Iglesias Fernández (eds). *Quality Assurance and Assessment Practices in Translation and Interpreting*. Hershey [PA]: IGI Global.

Translation quality in an error-annotated translation learner corpus

Andrea Wurm
Universität des Saarlandes
a.wurm@mx.uni-saarland.de

The aim of the paper is first to compare translation quality for groups of translator trainees with different backgrounds and grades and second to test if there are significant differences in translation quality and quantity before and after an intensive training unit.

Two seemingly obvious assumptions in translation competence (acquisition) (TC/TCA) are (i) that a good translation does not contain errors and (ii) that trainees make fewer errors in translation products after a certain amount of training (cf. the hypotheses in the PACTE TCA model, Orozco & Hurtado 2002: 388). The theoretical basis of the present paper is the model for TC and TCA and the importance of an acceptable (read: error-free) translation designed by the PACTE group (cf. e g. PACTE 2000, 2008, 2009, 2011), but aims at modelling (a) aspects of translation competence for groups of trainees and (b) timeline effects in TCA, such as (fewer) errors and good (better) solutions.

The analyses presented here are driven by the following research questions:
a) In what respects do trainee groups established on the basis of biographical data and grades differ in their translation quality?

- a1) Do trainees who spent time abroad make fewer errors and produce better solutions?
- a2) Are trainees who consume media products in their mother tongue better in terms of orthography, grammar and idiomatic expression?
- a3) Is there a significant difference in the number of good solutions produced by good, middle and low performance trainees?

b) In which way – concerning translation quality and quantity – do translator trainees respond to an intensive training unit?

- b1) Do individual translator trainees make fewer errors?
- b2) Do they come up with better solutions?
- b3) How are the results for the whole group?
- b4) Do trainees translate faster at the end of the training unit?
- b5) If yes: Are the results in terms of grades equal, better or worse than at the beginning?

To answer these questions, statistical methods shall be used. Various statistical analyses including univariate and multivariate (such as significance tests or supervised and unsupervised analyses) were applied in translation studies for different purposes, e. g. inter- or intra-annotator agreement, correlation of text difficulty and the acceptability of its translation and many other analyses (cf. e. g. PACTE 2011; Hvelplund 2011; Michael et al. 2014; Kunilovskaya 2015; Läubli 2014 and edited volumes in Corpus-based Translation Studies such as Oakes & Ji 2012 or De Sutter et al. 2017). Statistics can also be used in translation learner corpora research, which has become a dynamic field in the last few years. However, to my knowledge, no method for studying error annotations has been proposed, with the exception of Kunilovskaya (e.g. 2014) who is concerned with error frequencies normalized to text size, however without the use of statistical methods and Hansen (2006: 85, 137ff, 274ff) correlating error types and frequencies in translations made with or without time pressure or by mono- and bilingual students. Apart from the above-mentioned and other papers by Kunilovskaya and colleagues, there are very few studies reporting on error frequencies in translation learner corpora (Castagnoli et al. 2011; Sosnina 2005; Espunya 2014; Wurm 2013). Published corpus-based error analyses show a rather descriptive character of error frequencies and do not use statistical methods to explore the data further.

The data used in the present study was gained from a translation learner corpus with annotated negative and positive evaluation (KOPTE; Wurm 2016), covering parts of the bilingual, extra-linguistic, instrumental and strategic sub-competences in PACTE's componential TC model (e. g. 2009) with the categories of the evaluation scheme used (cf. also Hansen 2006: 113). The KOPTE corpus derives from a classroom setting and comprises target texts marked by the author. A total of 58 trainees handed in 968 target texts, of which 668 were marked using a fine-grained evaluation scheme (Wurm 2013). 43 trainees filled in a metadata questionnaire and delivered 532 evaluated target texts. For part (b) of the study, a subcorpus was extracted containing 468 target texts by 29 trainees who translated 10 or more evaluated texts. The corpus only reflects one trainer's assessment, but the detailed corpus annotation provides a means of exploring ways of analyzing data from translation learner corpora and thus may contribute to the design of subsequent studies. Moreover, some assumptions on TCA can be tested empirically on this dataset, bearing in mind the subjectivity of the trainer's evaluation.

An important part of the paper will be to test for significance between the evaluated translation quality for different trainee groups based on biographical features as well as between beginning and end of a training unit. To reach this goal, the trainer's evaluations in 7 categories (negative and positive each) and 75 more fine-grained criteria (44 negative and 31 positive) from KOPTE will be used in several aggregations. The absolute frequencies have been normalized over the respective token number and significance testing is conducted with these comparable values using chi-squared as well as Mann-Whitney U and t-test depending on data type (cf. e. g. Eichner 2010; Meindl 2011; Baur & Blasius 2014). Four different datasets shall be examined:

1) normalized frequencies of overall positive and negative evaluation
   a) for each trainee, total over all target texts
   b) for each target text for each trainee
2) normalized frequencies of evaluation categories/criteria in each target text
   a) for each trainee
   b) for the group of trainees.

Preliminary results based on boxplots of trainee groups with good, middle or low grades hint at a significant difference for the "low" group in positive evaluations, while a stay abroad does not seem to lead to better grades. In the presentation, I will situate the research questions in their TCA context and sketch a basic TCA model integrating empirical results. These will probably hint at translator competence profiles according to biographical features/grades as well as timeline effects in translator performance, esp. translation errors (and good solutions), in a population of translator trainees. Thus, for future work, the paper could highlight interesting phenomena in translation learner corpus research to be looked at more intently and/or tested under experimental conditions.

References

Baur, N. & Blasius, J. (eds). (2014). *Handbuch Methoden der empirischen Sozialforschung*. Wiesbaden: SpringerVS.

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

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

Eichner, G. (2010). *Datenanalyse mit R*. Vorlesungsskript. http://www.uni-giessen.de/cms/eichner, 21.09.13.

Espunya, A. (2014). The UPF Learner Translation Corpus as a resource for translator training. In *Language Resources and Evaluation* 48(1), 33-43.

Hansen, G. (2006). *Erfolgreich Übersetzen. Entdecken und Beheben von Störquellen*. Tübingen: Gunter Narr Verlag.

Hvelplund, K. T. (2011). *Allocation of Cognitive Resources in Translation: an eye-tracking and key-logging study*. Copenhagen. https://sites.google.com/site/ centretranslationinnovation/tpr-db-publications, 03.05.16.

Kunilovskaya, M. (2014). Error-tagging in Russion Learner Translator Corpus and its classroom applications. Presented at *didTRAD*, Barcelona, Spain (July 8, 2014). Extended paper on Academia.edu > Maria Kunilovskaya.

Kunilovskaya, M. (2015). How far do we agree on the quality of translation? In *English Studies at New Bulgarian University* 1(1), 18-31.

Läubli, S. [& Germann, U.] (2014). *Statistical Modelling of Human Translation Processes*. Unpublished Master's thesis, submitted to the School of Informatics, University of Edinburgh.

Meindl, C. (2011). *Methodik für Linguisten. Eine Einführung in Statistik und Versuchsplanung*. Tübingen: Narr.

Michael, E. B., Saner, L., Massaro, D., Bailey, B., de Terra, D., Messenger, S., Rhoad, K., Castle, S. & Campbell, S. (2014). Establishing Standards and Metrics for Translation: Experiments to Validate the Language Product Evaluation Tool. In J. W. Schwieter & A. Ferreira (eds). *The Development of Translation Competence: Theories and Methodologies from Psycholinguistics and Cognitive Science*. Newcastle upon Tyne: Cambridge Scholars Publishing, 169-200.

Oakes, M. P. & Ji, M. (eds). (2012). *Quantitative Methods in Corpus-Based Translation Studies. A practical guide to descriptive translation research*. Amsterdam & Philadelphia: John Benjamins.

Orozco, M. & Hurtado Albir, A. (2002). Measuring translation competence acquisition. *Meta* 47(3), 375-402. www.erudit.org/revue/meta/2002.

PACTE group (2000). Acquiring Translation Competence: Hypotheses and Methodological Problems in a Research Project. In A. Beeby, D. Ensinger & M. Presas (eds). *Investigating Translation*. Amsterdam: John Benjamins, 99-106.

PACTE group (2008). First results of a Translation Competence Experiment: 'Knowledge of Translation' and 'Efficacy of the Translation Process'. In J. Kearns (ed.) *Translator and Interpreter Training. Issues, Methods and Debates*. London: Continuum, 104-126.

PACTE group (2009). Results of the Validation of the PACTE translation competence Model: Acceptability and Decision making. In *Across Languages and Cultures* 10(2), 207-230.

PACTE group (2011). Results of the Validation of the PACTE Translation Competence Model: Translation Project and Dynamic Translation Index. In S. O'Brien (ed.) *Cognitive Explorations of Translation*. IATIS Yearbook 2010. London: Continuum, 30-53.

Sosnina, E. (2005). Russian Translation Learner Corpus: The First Insights. In *Proceedings of the 6th International Scientific Conference "Interactive Systems: Problems of Human-computer Interaction"* (Vol. 1). Ulyanovsk: UlSTU, 60-61.

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.uni-saarland.de/fachrichtung/lst/staff/andrea-wurm.htm.

Quantitative corpus approaches in comparative literary translation analyses

Jitka Zehnalová
Palacký University Olomouc
jitka.zehnalova@upol.cz

The contribution is a part of a research project dealing with translation strategies used by contemporary literary translators from English into Czech. Translating is conceptualised as a social activity and translation strategies as decision processes that are governed by norms (Toury 1995) and that can be investigated by combining translational and sociological methods. The contribution focuses on the translational part of the project, more specifically on the methodology of the comparative analyses of source texts (STs) and target texts (TTs). The analyses are designed to discover "regularities in the observable results of a particular kind of behaviour, assumed to have been governed by norms" (Toury 1999: 15). They are conducted on a set of literary translations from English into Czech published in the time period 2000-2016, and their STs.

To provide a framework for the present contribution, the research results obtained so far are first briefly introduced: the methods of creating the basic set of TTs (N=15,381), the methods of compiling the representative selection set of TTs (n=854), the parameters used for the targeted selection of translators and TTs out of the representative selection set, and a pilot set of TTs and their translators. In the current research phase, this pilot set is employed to evaluate the suitability of corpus-based methods of literary text analysis for the given research project and its aims. The substantial part of the contribution discusses the possibilities of applying the method of comparing the *thematic concentration* of STs and TTs via the so called *h-point*. The methodology is based on a research monograph dealing with quantitative corpus based methods of text analysis within Literature Studies (Změlík 2015) and it has not yet been applied in the context of bi-lingual translational analyses. The method consists of subsequent steps, which, ordered according to their expected effectiveness (perceived as the relevance of the obtained results compared with the time and difficulties involved) from the most to the least effective, include:

1. Identifying of thematically marked vocabulary: An appraisal of the *thematic concentration* of the ST and the TT via the so called *h-point*, "which is a position separating two neighbouring areas of word distribution: it indicates the relation between the order of a word according to its frequency *r* and its frequency f_r. […] h-point is thus a position for which r = f(r)" (Změlík 2015: 124). The items under analysis are autosemantic words above h-point, so called *thematic words* used in the ST and the TT. While the h-point is a relative divide rather than a sharp dividing line between the thematic and non-thematic parts of the lexicon, it still considerably contributes to identifying thematically marked vocabulary that signals the thematic orientation of the text.

2. Identifying and statistical evaluation of differences between STs and TTs in terms of the *thematic words* used: Several Czech studies dealing with the subject of the so called *thematic concentration of text* (Glogarová & Čech 2013; Popescu et al. 2014; David et al. 2013; Čech 2016) have demonstrated the usefulness of this procedure in monolingual analyses. In the context of translational analyses, its usefulness has to be tested. The findings are expected to uncover differences between STs and TTs in terms of thematically marked lexicons and thus to point to translation strategies.

3. Creating of frequency profiles of the thematically marked lexicons: The previous results can be complemented by another statistical analysis, which enables modelling on the basis of frequency features. The output of this analysis is frequency profiles of the thematically marked lexicons; these can be compared and the results interpreted.

4. Comparing of the frequency markedness of thematic words: This is a comparison of the frequency of ST thematic words with their frequency in referential source language corpora and a comparison of the frequency of TT thematic words with their frequency in referential target language corpora.

5. Comparing of general semantic classes: Based on the data obtained, the distribution of the identified thematically marked lexical units into general semantic classes such as movement, subject, object, time, space (in the narrow sense of setting), features of the subject, space (broader spatial horizon) can be conducted, aiming at establishing the extent and significance of differences between STs and TTs in this respect, i.e. in the number and lexical representation of semantic classes. This step is suggested as a possible follow-up procedure.

To carry out this type of analysis, the online available parallel corpus *Intercorp*, a part of the *Czech National Corpus*, is utilised, as well as small purpose-built corpora (size at the moment: 139,796 positions) of English STs and their translations into Czech that were compiled from a pilot set of texts via the corpus manager *Sketch Engine*.

The research seeks to answer these questions:

a) Is the suggested method applicable/useful for comparative translational analyses?
b) Does it contribute to identification of translation strategies?
c) What is an adequate way of interpreting the results of quantitative analyses transformed into statistical models?
d) Based on the conducted research, is it reasonable to increase the use of the existing parallel corpora (*Intercorp*), or even to compile other small purpose-built corpora?

The contribution is designed as a pilot study. That is why the obtained data and results are looked upon as probes meant to prove/disprove the usefulness of the suggested method and are offered for critical discussion and evaluation.

References

Čech, R. (2016). *Tematická koncentrace textu v češtině*. Praque: Ústav formální a aplikované lingvistiky.

David, J., Čech, R., Davidová-Glogarová, J., Radková, L. & Šústková, H. (2013). *Slovo a text v historickém kontextu: perspektivy historickosémantické analýzy jazyka*. Brno: Host.

Davidová-Glogarová, J. & Čech, R. (2013). Tematická koncentrace textu – některé aspekty autorského stylu Ladislava Jehličky. *Naše řeč* 96, 234-245.

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

Mikhailov, M. & Cooper, R. (2016). *Corpus Linguistics for Translation and Contrastive Studies: A guide for research*. New York: Routledge.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam & Philadelphia: John Benjamins.

Toury, G. (1999). A Handful of Paragraphs on 'Translation' and 'Norms'. In Ch. Schäffner (ed.) *Translation and Norms. Current issues in language and society*. Clevedon & Philadelphia: Multilingual Matters, 9-31.

Vogel, C., Lynch, G., Moreau E., Mamani Sanchez, L. & Ritchie, P. (2013). Found in translation. Computational discovery of translation effects. *Translation Spaces* 2, 81-104.

Změlík, R. (2015). *Kvantitativně-korpusová analýza a literární věda: model a realizace autorského korpusu a slovníku Jana Čepa v kontextu zahraniční a české autorské lexikografie*. Olomouc: Palacký University Olomouc.

On-line Sources

http://www.korpus.cz/intercorp/
https://www.sketchengine.co.uk/

A corpus-based comparative study on football texts and their translation strategies

Tianqi Zhang
Universitat Autònoma de Barcelona
tianqi.zhang@e-campus.uab.cat

As the most important sport worldwide, football undoubtedly plays a decisive role in the process of globalization. The language of football, in turn, has become a fundamental linguistic bridge which connects the football sport and its community. The journalistic language of football was born from the oral narration and written commentary on football matches by sports journalists with the main purpose of representing the game with the help of linguistic resources in the best way for football fans who are not able to watch the match live. Throughout the development of modern football of the last century, the language of football has undergone its own evolution and displays some unique stylistic features, such as the use of metaphors, high occurrence of synonyms and technical terms which belong to the sport itself (Hernández Alonso 2003). The language of football has been gradually accepted by the public. In countries with a long tradition of football such as England, Germany or Spain, the languages of football often occupy an influential position as a sublanguage within the general language.

Thanks to the Chinese government's emphasis on football as a national sport during the last two decades there have been large numbers of international exchanges of football professionals between China and overseas. In March 2015, an Overall Reform for Chinese Football Reform and Development was issued by the General Office of the State Council of People’s Republic of China. Since then, the demand for translation of football texts has increased substantially. Likewise, requirements for the quality of translated texts have also increased.

Consequently, high-quality translation has become the first need in the football sector in order to achieve effective and efficient communication between different language speakers. However, due to the lack of systematic study of football languages in Chinese and other non-English languages, the quality of the translated football texts is still poor from the perspective of the football industry. As a matter of fact, the translation of football texts is mostly carried out either by sports journalists themselves using English as a lingua franca, or by translators without sufficient football knowledge. In neither case is the translation quality satisfactory. As a result, a large number of synonyms can be detected due to different translations of the same term from the source language, causing confusion between the new concepts for the readers of the target language. From the academic perspective, the language of football is always considered as a less specialized language compared to other domains such as medicine or physics, and thus it lies at the border of specialized and common languages (Hurtado 2001: 60). To date, the translation of football texts has not been investigated systematically with translatological methods (Zhang & Aguilar-Amat 2017). Above all, it is necessary to establish an optimized classification of football texts according to their different functions and levels of specialization.

In this article we observe the journalistic language of football in Spanish and Chinese by using the online corpus tool Sketch Engine. We compared the list of frequent words, the use of metaphors, the appearance of synonyms and football terms in a comparable corpus of the two languages. We have built our own comparable corpus due to the lack of direct translation of football texts between Chinese and Spanish. According to the proposal of Franquesa (2008: 35) on specialized texts, we have classified the football texts into three types, i.e., informative, commentary and directive texts. We chose the chronicles of football matches as informative-commentary texts. The criteria for the selection of texts are based on several parameters, such as the same match on the same date, the narration of the same plays by the same players, and chronicles written by sports journalists in their mother languages. The chronicles should contain as many football terms as possible.

With the obtained results after the analysis of the corpus, we arrived at the conclusion that the football language of Chinese and Spanish share several similarities, and we have also confirmed the hypothesis of the homogeneity of the football language. On the one hand, football language features many anglicisms due to the use of football terms in English as a lingua franca among sports journalists and the frequent appearance of synonyms for the same concept. On the other hand, in both Chinese and Spanish football languages, journalists use numerous metaphors to achieve a more invitational tone for the reader. Taking these phenomena into account, we have proposed several parameters with respect to the translation strategy of football texts, such as the rejection of loanwords when the term can be already found in the target language. Likewise, special attention must be paid to the translation of metaphors into the football language to maintain their expressive function. Translation equivalence should be sought as far as possible in the target culture, and the vividness of football language should not be lost.

References

Bertaccini, F., Massari, M. & Castagnoli, S. (2010). Synonymy and variation in the domain of digital terrestrial television. *Terminology in everyday life* 13-11.

Castellví, M. T. C. & Cormier, M. C. (1998). La terminologie : théorie, méthode et applications. Ottawa: Presses de l'Université d'Ottawa.

Franquesa i Bonet, E. F. (2011). La terminologia: un mirall del món. Editorial UOC.

General Office of the State Council of People's Republic of China. (2015). *Overall Reform for Chinese Football Reform and Development*. Guo Ban Fa No.11. 2015-03-08.

Hernández Alonso, N. (2003). *El lenguaje de las crónicas deportivas*. Madrid: Ediciones Cátedra.

Newmark, P. (1988). *A textbook of translation* (Vol. 66). New York: Prentice hall.

Zhang, T. & Aguilar-Amat, A. (2017). China y medios de comunicación. La traducción española/chino de términos futbolísticos en los medios de comunicación escrita. *Communication Papers*, 6(11), 27-49.