Analysing complex contrastive data

Jenny Ström Herold[1], Magnus Levin[1], Signe Oksefjell Ebeling[2], Anna Čermáková[3]

[1]Linnaeus University (Sweden), [2]University of Oslo (Norway), [3]Charles University, Prague (Czech Republic)

1. Introduction

This collection of papers is the result of the contrastive pre-conference workshop at the 41st ICAME[1] conference, *Language and Linguistics in a Complex World: Data, Interdisciplinarity, Transfer, and the Next Generation*, held in a Covid-19-safe distance format at Heidelberg University, Germany, May 20–23, 2020. ICAME41 had an ambitious goal of taking "corpus linguistics out of its comfort zone" and "to emphasise that language is the crucial social and cultural factor in human interaction".[2] The theme of the workshop, *Crossing the Borders: Complex Contrastive Data and the Next Generation*, tied in closely with the focus of the main conference. The aim was to expand the previous focus of contrastive corpus-based studies from bilingual comparisons of mostly lexicogrammatical features to include new types of synchronic or diachronic corpus data, new language pairs – in particular going beyond the traditional two-language perspective –, new areas of investigation such as semantics, pragmatics and phraseology combined with methods and interdisciplinary approaches.

The workshop contributors responded to the challenge and explored complex data in multi-lingual settings, most studies using comparable data and some also investigating parallel/translation corpora. Thus, this publication offers contributions which, taken together, involve seven different languages from a range of language families – Czech, Finnish, French, German, Norwegian, Spanish and Swedish –, and which are contrasted with English. Some studies take a three-language approach, or more, and some focus on areas that have traditionally been under-investigated in contrastive and translation studies, such as punctuation and phraseological patterns. Yet, there are contributions which take a diachronic approach, but also those which use synchronic corpus data from "innovative" genres that have received little attention in contrastive studies. What most papers have in common, though, is that they are based on relatively small – both parallel and comparable – corpora compared to large-scale present-day mono-lingual corpora. But, as can be seen from this overview, this has not restricted the inventiveness or curiosity of the researchers represented in this volume. The smaller datasets also allow the insightful qualitative analyses typical of such studies.

Carefully sampled small-scale contrastive data, as partly seen in the present volume, is a sound starting point for qualitative analyses of differences and commonalities between languages. The restricted size of such corpora may, however, be criticized due to limited

[1] International Computer Archive of Modern and Medieval English (http://clu.uni.no/icame/).
[2] https://icame41.as.uni-heidelberg.de/theme/

CROSSING THE BORDERS: ANALYSING COMPLEX CONTRASTIVE DATA. Edited by Anna Čermáková, Signe Oksefjell Ebeling, Magnus Levin and Jenny Ström Herold. *BeLLS* Vol 11, No 1 (2021), DOI: 10.15845/bells.v11i1.3434.

generalizability. Such considerations do not only pertain to contrastive linguistics involving English, but also mono-lingual English corpus linguistics. Still, there are, in Mair's (2006) words, clear advantages of traditional "small and tidy" corpora when comparing with the shortcomings of "big and messy" corpora. These considerations from mono-lingual corpus studies are no less pertinent in the area of contrastive corpus-based linguistics. When performing in-depth qualitative analyses on multi- or mono-lingual data, the smallness and tidiness of the samples is beneficial. Restricted data size allows researchers to work on the material "under controlled conditions" and crucially, for contrastive studies, to ensure data comparability. For example, with a small and tidy corpus, it is easier to keep an overview of what is included in the data both regarding content and structure. In the present volume, the studies range from traditional small and tidy corpora (e.g., ENPC and OMC) and newer small-scale corpora (e.g., CLANES and LEGS) to large-corpora, not usually seen in contrastive studies (e.g., CLMET). Thus, the wide range of corpora used here indicates that one size does not fit all (Egbert *et al*., 2020: 4), but instead the choice of corpus largely depends on the research questions. The present volume thus fulfils our aim of expanding the traditional horizons of contrastive corpus-based studies.

2. Structure of this volume and presentation of contributions

The ten contributions in this volume all contrast English with at least one other language, using both standard corpora and more recently compiled specialized corpora. No fewer than 12 corpora are investigated in the present volume, including both multi-lingual and mono-lingual corpora:

- *Multi-lingual corpora*
    - Controlled LANguage English Spanish (CLANES); see Rabadán *et al*.
    - English-Norwegian Match Report Corpus (ENMaRC); see Ebeling
    - English-Norwegian Parallel Corpus(+) (ENPC and ENPC+); see Ebeling; Egan; Hasselgård
    - Linnaeus University English-German-Swedish corpus (LEGS); see Levin and Ström Herold; Ström Herold *et al*.
    - Multilingual Parallel Corpus (MPC); see Viberg
    - Oslo Multilingual Corpus (OMC); see Egan
- *Mono-lingual English corpora*
    - British National Corpus (BNC); see Čermáková and Malá; Šebestová
    - Corpus of Late Modern English Texts (CLMET); see Krielke
    - Royal Society Corpus (RSC); see Krielke
- *Mono-lingual corpora of other languages*
    - Czech National Corpus (CNC); see Čermáková and Malá; Šebestová

- Deutsches Textarchiv (DTA); see Krielke
- Savokorpus (Finnish); see Čermáková and Malá

The contributions in this volume are presented below. In order of appearance, these include (i) studies on lexical searches that enable explorations of phraseological patterns, broadly construed, (ii) papers that primarily have a syntactic focus and (iii) studies, in which contrastive data is used to analyze textual and discourse phenomena.

Signe Oksefjell Ebeling explores the English and Norwegian cognate nouns and verbs HOPE/HÅP(E) and their collocations and phraseological patterns. The material combines online football match reports from ENMaRC and fiction from ENPC+. The findings indicate both cross-linguistic and genre-specific differences. So, for instance, the nouns are more frequent in match reports in both languages, while the verbs predominate in fiction. This finding is in accordance with previous findings on noun and verb usage in news and fiction. A notable result is that the English lemmas more often occur in negative contexts, as for example with 'hope' being *extinguished*, *quashed* or *killed off*, than their Norwegian counterparts. The comparison of two genres across two languages thus sheds new light on what features are genre-specific and what features are language-specific.

Denisa Šebestová's contribution compares the phraseology connected to the English preposition *in* and its Czech equivalent *v* ('in'). These prepositions are highly frequent in the investigated material, the BNC and the CNC. The findings indicate considerable similarities between the two languages, in spite of their typological differences. Among the cross-linguistically frequent categories identified in the corpora, there are adverbials such as *in this respect* and *v tomto ohledu* ('in this respect'), complex prepositions such as *in front of* and *v rámci* NP ('within NP') and various pragmatic hedging patterns (*in a sense*). Some typological preferences also emerge: the more analytic English language contains more complex prepositions and conjunctions than the more synthetic Czech. The findings produced can be applicable in teaching practice. Foreign-language learners have been found to have difficulties acquiring a large and varied repository of (semi-)fixed phrases in the target language, and such contrastive data can therefore provide valuable input to learners.

Thomas Egan presents the results of a tri-lingual study of TELL predications in English, Norwegian and French, targeting the cognate verbs English *tell* and Norwegian *fortelle* and French renditions such as *dire* ('say'). The data was collected from the ENPC and the OMC. The results show that *tell* and *fortelle* in English and Norwegian original texts are very different in their lexico-grammatical behaviour. *Tell* is also more than four times as common and occurs with a greater syntactic variety of THEMEs than *fortelle*. As for translations, tokens with NP THEMEs are most often translated congruently, both in the Norwegian → English and the English → Norwegian direction. One striking observation is that Norwegian translations are inclined to employ the more neutral reporting verb *si* ('say'), most likely because *si*, unlike its English cognate *say*, easily combines with indirect objects. The results from French translations suggest that French is more similar to Norwegian than English, one reason being that the verb *dire*, like Norwegian *si*, can take an indirect object, which makes it an appropriate correspondent of many English ditransitive *tell* predications.

Åke Viberg's contribution concerns a comparison of the Swedish particles *upp* ('up') and *ner* ('down') indicating the endpoint of motion across four languages – the Germanic English and German, the Romance French and the Finno-Ugric Finnish. The comparisons show that there are both differences related to inter-family features but also to intra-family preferences. Using the MPC consisting of Swedish novels translated into the four languages, the study illustrates the differences between satellite-framed languages, where the path is expressed in satellites outside the verb (such as English *go up* or Swedish *gå upp*) and verb-

framed languages, where the path is encoded in the verb (as in French *monter* 'move-up'). In the German and Finnish translations, the particle is often rendered as zero while the positional change is indicated by case. In these two satellite-framed languages, in contrast to Swedish and English, verticality is not expressed, which suggests that there are differences within this set of languages based on morpho-syntactic differences.

Marie-Pauline Krielke's paper is a diachronic English-German study investigating the changing levels of grammatical complexity from the 17th to the 19th centuries. Relativizers (relative clauses) are here the chosen proxy. The study includes a cross-register comparison of general and scientific language, based on comparable texts from three corpora: for English, the RSC and the CLMET, and for German, the DTA. The main hypothesis is that scientific texts, over time, become grammatically less complex, using fewer relative clauses, as compared to general texts. This is found to hold true, but it is a development that pertains also to general language – in both English and German. However, in German scientific language, grammatical complexity is shown to decrease much later than in English. The fact that the German decrease does not happen until the second half of the 18th century may be due to several factors. One of these factors seems to be the longstanding Latin influence on German scientific writing.

Using the English-German-Swedish LEGS corpus, **Magnus Levin** and **Jenny Ström Herold** investigate the use of round brackets in originals and translations. Brackets are found to be most frequent in English non-fiction and the least frequent in Swedish. English translators introduce the most changes by adding or omitting brackets, or by changing punctuation marks. Swedish translators, in contrast, are the most conservative and introduce less changes than either English or German translators, a result which seems to reflect a status difference in the languages. Commas or zero punctuation are, apart from brackets, the most frequent translation correspondences in all translation directions. When translators introduce brackets, these often involve the addition of short synonyms, irrespective of translation direction. The intricate structure of the corpus with three original languages and six different translation directions enables the separation of language-specific preferences and translation trends.

Hilde Hasselgård's paper compares 'noun + preposition' sequences in English and Norwegian fiction texts in the ENPC. Postmodifiers turn out to be the most frequent function in both languages, followed by adverbial. The preference for postmodifiers is even stronger in English than in Norwegian. These findings suggest that English prefers more phrasal modes of expression with Norwegian being more clausal in nature. Regarding the translations, Hasselgård finds that adverbials are more often translated congruently than postmodifiers, and that this tendency is particularly prevalent in translations from English into Norwegian. The reason for this specific lack of congruence is the English preposition *of*, which lacks a direct correspondent in Norwegian. Translations from Norwegian, in contrast, do not encounter the issue of dealing with *of*, and are therefore more congruent. The paper illustrates that relying on tag sequences is a bottom-up approach that can be used by researchers to retrieve patterns that would not otherwise be identified.

The contribution by **Jenny Ström Herold**, **Magnus Levin** and **Jukka Tyrkkö** deals with acronyms in English, German and Swedish from the LEGS corpus. More specifically, it targets translation strategies employed by German and Swedish translators when encountering universal (*DNA*) and culture-specific (*SAT*) acronyms in English original texts. Here, the contrastive perspective holds mainly between the German and Swedish target texts, the main parts of the study, however, being geared towards the translation perspective. Due to morphosyntactic differences, English acronym premodifiers often merge into hyphenated compounds in German translations, but less frequently so in Swedish translations. Swedish translators are more inclined to using prepositional phrases as correspondences. When introducing acronyms, German translators explain and elaborate more on English acronyms than Swedish translators and they do so preferably in the German language. Swedish translators

instead use English to a greater extent, suggesting that Swedish readers are expected to have better knowledge of English than German readers. Overall, the study reveals a range of explanation strategies where translators elaborate on English acronyms by, e.g., adding a spelt-out version of the English acronym.

Anna Čermáková and **Markéta Malá**'s contrastive study concerns eye-behaviour in fictional speech. It is based on data from three typologically different languages: English, Czech and Finnish. Children's fiction in original is analysed, drawing on three comparable corpora – the BNC, the CNC (SYN-7) and the Savokorpus –, and the paper explores the distribution and use of the 'eye' lemmas EYE, OKO and SILMÄ. Both grammatical and narrative functions are discussed across the languages. In terms of syntactic encoding, the study shows that EYE in English is more strongly associated with the subject/agent role than OKO in Czech and SILMÄ in Finnish. Czech and Finnish preferably introduce the 'eye phrase' through an adverbial phrase expressing location. As for narrative functions, the three languages behave similarly: eye-behaviour descriptions support the speech by highlighting the content or the manner of speaking. The study thus suggests that 'eyes' are a vital part of the narrative in all languages, the examined languages sharing various communicative and interpersonal functions, but that the grammatical behaviour may differ depending on language type.

The contribution by **Rosa Rabadán**, **Noelia Ramón** and **Hugo Sanjurjo-González** addresses the more technical side of annotating a parallel corpus. The authors present a model for pragmatic annotation of their comparable English-Spanish CLANES corpus comprising informational-promotional and instructive texts about gourmet foods and drinks. The pragmatic annotation involves a combination of the semantic annotation scheme, the UCREL Semantic Analysis System, together with part-of-speech tagging. The paper identifies seven different pragmatic functions such as <DIRECT> (e.g., *remove the pan from the stove*) and <PRAISE> (e.g., *truly lovely cheese*). The trials show promising results regarding accuracy but a number of challenges are also identified. For instance, the segmentation of the text was sometimes problematic due to the lack of punctuation in headings, and a lot of hands-on labour was needed for corrections, partly because the part-of-speech tagset differs between English and Spanish. The ultimate aim of the authors' ongoing annotation project is to provide support to authors writing about food and drinks.

Acknowledgements

We would like to express our gratitude to all contributors to this volume for their submissions, revisions and excellent cooperation. We also gratefully acknowledge the key contributions made by the anonymous peer reviewers for their timely and insightful comments. A professional peer-review process is key for any high-quality academic publication, and the peer reviewers are often unsung heroes in the process. Our thanks are also due to the organizers of the ICAME41 conference at Heidelberg who were able to organize a conference during the Covid-19 pandemic. Finally, we extend our thanks to the general editors of *Bergen Language and Linguistics Studies*, and Dr. Lidun Hareide in particular, for enthusiastically accepting this volume in their series, and to Tormod Eismann Strømme at Bergen University Library for technical support.

References

Egbert, J., Larsson, T. and Biber, D. 2020. *Doing Linguistics with a Corpus. Methodological Considerations for the Everyday User*. Cambridge: CUP.

Mair, C. 2006. Tracking Ongoing Grammatical Change and Recent Diversification in Present-day Standard English: The Complementary Role of Small and Large Corpora. In *The Changing Face of Corpus Linguistics*, A. Renouf and A. Kehoe (eds), 355–376. Leiden: Brill/Rodopi.