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.

Hope for the future: An analysis of HOPE/HÅP(E) across genres and languages

Signe Oksefjell Ebeling

University of Oslo (Norway)

This article reports on a contrastive study of the cognate nouns and verbs HOPE and HÅP(E) that investigates their lexico-grammatical conditions of use in English vs. Norwegian fiction texts and football match reports. The complex dataset consists of material from a parallel corpus of fiction texts and a comparable corpus of football match reports. An interesting finding is that the verb use outnumbers the noun use in the fiction texts, whereas the noun use outnumbers the verb use in the match reports in both languages. Moreover, the analysis of the lemmas suggests that they have similar potential of use but with slightly different preferences, both across the genres and languages. It is also suggested that the English lemmas are more consistently used in negative contexts than the Norwegian ones. Finally, the method of combining data from two different types of contrastive corpora proved fruitful, as the results become more robust.

Keywords: cognates, comparable corpus, bidirectional parallel corpus, fiction, football match reports, English/Norwegian

1. Introduction and aims

This article presents a contrastive analysis of the cognates HOPE and HÅP(E) through their use in two languages (English and Norwegian) and two genres (fiction and football match reports). In a previous contrastive study of English and Norwegian football match reports it was found that the cognate nouns HOPE and HÅP featured as keywords in texts reporting on defeat (Ebeling, 2019). The reason for this frequent use of HOPE and HÅP in the defeat section of the English-Norwegian Match Report Corpus (ENMaRC) was attributed to the items' frequent use in contexts where hopes are dashed, as in examples (1) and (2).

(1) However those *hopes* were dashed on 55 minutes when the Gunners added a second. (ENMaRC/CPFC)[1]

(2) Det tente et ørlite *håp* som ble knust desto mer brutalt fem minutter etter. (ENMaRC/VFK)
Lit.: That lit a tiny hope that was dashed even more brutally five minutes later

[1] The ENMaRC corpus text code is the same as the official acronyms of the clubs (CPFC = Crystal Palace Football Club). See Ebeling (2021) for an overview of clubs and acronyms in the ENMaRC.

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.v11i1.3436.

It was speculated that the use attested in (1) and (2) may be more typical of match reports than of language in general. The fact that HOPE and HÅP feature in reports describing defeats suggests that words may be coloured by their immediate context to the effect that they take on a meaning that is the opposite of what might otherwise be the case (in their typical contexts in other genres). Therefore, part of this study aims to find out to what extent this (negative) use of the nouns HOPE and HÅP and their verb counterparts (HOPE and HÅPE) is overrepresented in the genre of football match reports, or whether it may be seen to extend to other genres.

As mentioned, the second genre under scrutiny here is fiction, and the reason for this choice is twofold. First, it was deemed necessary to investigate the use of the lemmas in a genre that is clearly distinct from written match reports.[2] Second, from a contrastive perspective, it was deemed necessary to objectively determine the degree of equivalence between the English and Norwegian lemmas on the basis of bidirectional translation data to be certain that we compare like with like. In other words, an English-Norwegian bidirectional corpus had to be consulted, and the only corpus containing enough data from one relatively homogeneous genre is the English-Norwegian Parallel Corpus+ (ENPC+) of fiction texts.

Drawing on data from the ENMaRC and ENPC+, this study seeks to dig deeper into the lemmas HOPE and HÅP(E) by contrasting their lexico-grammatical conditions of use across languages and genres. Preliminary scrutiny of concordance lines suggested that the lemmas have the potential to feature in a number of different (phraseological) contexts in both languages. However, when compared to fiction, match reports seem to make a narrower selection from the full potential of contexts in which the lemmas may be used (cf. Stubbs, 1996: 89).

Based on these observations and previous findings, the study seeks answers to the following questions:

- To what extent are the lemmas used similarly in English and Norwegian?
- To what extent are the lemmas used similarly in match reports and fiction?
- To what extent does the use of different types of "contrastive" corpora contribute to our cross-linguistic knowledge of the lemmas?

The aim is to provide new insights into the actual use and lexico-grammatical features of these lemmas, not only from a cross-linguistic perspective but from a cross-linguistic genre perspective. This ties in with a recent trend in contrastive studies, in which more attention is given to cross-linguistic variation across genres or registers (see e.g. Dupont and Zufferey, 2017; Lefer and Vogeleer, 2014; Neumann, 2014; Teich, 2003). Moreover, the study addresses potential benefits of using both comparable and (bidirectional) translation corpora to widen the horizons of contrastive studies.

The study starts with a general description of the rather complex data under investigation by introducing the corpora used in Section 2. An outline of the contrastive approach and method applied is offered in Section 3, including an account of how the data were extracted and an overview of the material used in the analysis. A cross-linguistic, cross-genre analysis of the actual uses of the lemmas is carried out in Section 4, followed by a discussion of some of the findings in Section 5. Section 6 revisits the research questions and offers some concluding remarks and suggestions for future studies.

[2] Cf. Biber's (1993: 336) multidimensional analysis, where fiction and press reportage, of which match reports can be seen as a sub-register, are shown to differ according to several linguistic features.

2. Corpora

As mentioned in Section 1, the material for the present study is extracted from the English-Norwegian Parallel Corpus+ (ENPC+) and the English-Norwegian Match Report Corpus (ENMaRC). The ENPC+ is a bidirectional translation corpus of published fiction and its structure and potential are illustrated in Figure 1 (Johansson and Hofland, 1994). It contains 39 fiction texts originally written in each of the languages with their translations into the other. The texts were published in the period from 1980 to 2012 and include both full-length novels (eight in English and nine in Norwegian) and extracts of 12,000 to 15,000 words (31 in English and 30 in Norwegian). In total, the ENPC+ contains ca. 5.3 million words, i.e. roughly 1.3 million words in each of the sub-corpora: English originals, Norwegian originals, English translations, Norwegian translations. For a more detailed description of the ENPC+, see Ebeling and Ebeling (2013).

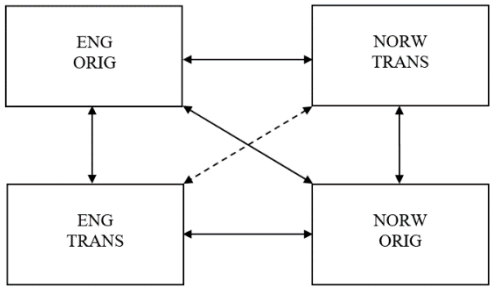

Figure 1. The bidirectional structure of the ENPC+.

This corpus structure enables contrastive studies of a comparable nature, using material from the original texts only, as well as of a parallel nature, using material from the original and their aligned translated texts in both directions. From a translation studies perspective, the potential of comparing original and translated texts in the same language is also a valuable feature of this corpus structure (see e.g. Ebeling and Ebeling, 2017; Ebeling, forthcoming), although not relevant to the present study.

The English-Norwegian Match Report Corpus is a comparable corpus of online written football match reports from the English Premier League and the Norwegian *Eliteserie*. It is comparable according to Johansson (2007: 9), in the sense that it contains original texts in two languages matched by criteria such as genre, time of publication, etc. (see also Ebeling and Ebeling 2020). Its structure is illustrated in Figure 2, corresponding to the boxes connected by the slant solid double arrow in Figure 1.

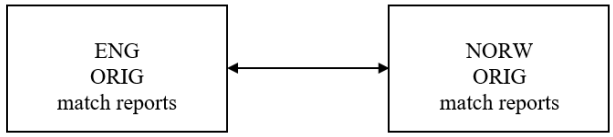

Figure 2. The structure of the ENMaRC.

The match reports are written by the respective football clubs' own journalists and published online on the clubs' web pages immediately after each match. The ENMaRC contains match reports from two seasons, 2016–17 and 2017–18 in the case of the Premier League and 2017 and 2018 in the case of the *Eliteserie* (ES).[3] The Premier League part of the corpus contains

[3] The football season in England runs from August until May (hence 2016–2017 and 2017–2018) and the season in Norway runs from March until November (hence 2017 and 2018).

reports from 23 teams and amounts to approximately 990,000 words, while the *Eliteserie* part contains reports from 18 teams, amounting to around 315,000 words. Although there are some challenges relating to corpus size when comparing the use of HOPE and HÅP(E) in fiction and match reports, these will be kept to a minimum through the use of normalised frequencies, some (mainly descriptive) statistics and qualitative case studies. Another potential challenge relating to corpus comparability is the time period covered in the ENPC+ vs. the ENMaRC. However, it is not believed that the use of these lemmas has changed much since the earliest texts in the ENPC+ (1980s) to the most recent texts in the ENMaRC (2018).

3. Contrastive approach, method and material

With both a comparable corpus (ENMaRC) and a bidirectional parallel corpus (ENPC+) at hand a generally sound contrastive approach is ensured and the contrastive corpus model ensuing from the combination of the two can be illustrated as in Figure 3.

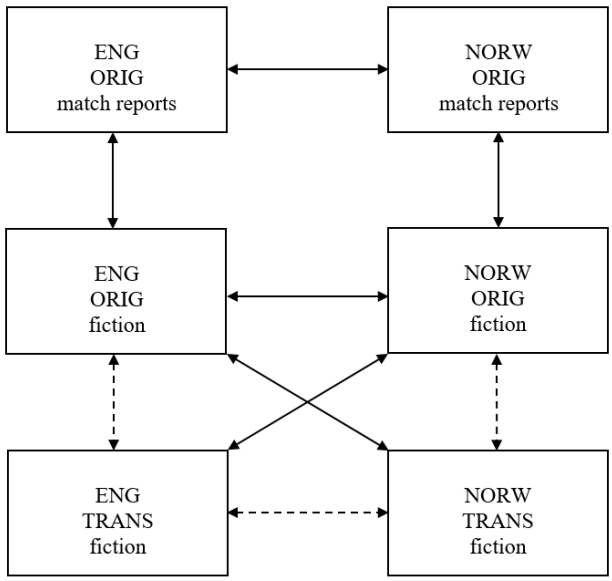

Figure 3. The two-genre comparable-cum-bidirectional corpus model.

The structure and potential of the model can be summed up as a two-genre comparable-cum-bidirectional corpus model. For the purpose of this study, the bidirectional fiction part is mainly used to objectively establish the comparability of the items compared by assessing the items' intertranslatability in a measure of Mutual Correspondence (Altenberg, 1999). Although HOPE and HÅP(E) are etymologically cognate,[4] and as such fulfil the criterion of the presence of the comparability criterion of a perceived similarity as outlined by Chesterman (1998: 54), their comparability is further strengthened by a Mutual Correspondence of a staggering 95% for the verbs HOPE and HÅPE and an almost equally staggering 91.3% for the nouns HOPE and HÅP in the ENPC+.[5] This demonstrates that the lemmas are very good cross-linguistic matches of each other and they can safely serve as the starting point of a contrastive analysis. Typical examples are shown in (3) from English into Norwegian and in (4) from Norwegian into English.

[4] From Middle Low German and Middle Dutch *hope* (*Oxford English Dictionary* and *Det Norske Akademis ordbok*).

[5] Mutual Correspondence refers to "the frequency with which different (grammatical, semantic and lexical) expressions are translated into each other", ranging from "0% (no correspondence) to 100% (full correspondence)" (Altenberg, 1999: 254).

(3) Long after all *hope* had gone Stanton stood there and waited for something to happen... [ENPC+/MoAl1E][6]
Lenge etter at alt *håp* var ute, sto Stanton der og ventet på at noe skulle skje ... [ENPC+/MoAl1TN]

(4) — La oss *håpe* at snøen dekker ham til før noen oppdager at han ligger der. [AnHo2N]
"Let's *hope* the snow will cover him before anybody sees him. [AnHo2TE]

In the cross-linguistic, cross-genre analysis proper, the ENPC+ will not be used to its full potential, and the contrastive analysis will from now on be based on the comparable texts only: fiction and match reports originally written in English and Norwegian (cf. the top four boxes in Figure 3). In the following, the steps taken in the analysis will be described, focusing on the lemmas' phraseological potential in the two languages and genres.

The first step was to search for all forms of the lemmas using the ENPC+ search interface for the fiction texts[7] and AntConc (Anthony, 2019) for the match reports. As the corpora are not part-of-speech tagged, manual disambiguation of noun and verb uses had to be performed on the full set of search results (raw numbers): 375 and 112 for the Norwegian forms (håp|håpet|håper|håpte|håpa)[8] in the ENPC+ and ENMaRC, respectively, and 450 and 324 for the English forms (hope|hopes|hoped|hoping)[9].

Table 1 shows the number of occurrences of all noun and verb forms of HOPE and HÅP(E) in the four sub-corpora, both in terms of raw frequencies and normalised frequencies per 100,000 words.

Table 1. Number of occurrences of HOPE and HÅP(E) in the ENPC+ and ENMaRC.

	ENPC+ Occ. per 100,000 words (Raw freq.)			**ENMaRC** Occ. per 100,000 words (Raw freq.)		
Word forms	**Noun**	**Verb**	**TOTAL**	**Noun**	**Verb**	**TOTAL**
hope|hopes|hoped|hoping	6.2 (83)	27.3 (367)	33.5 (450)	23.7 (235)	8.9 (88)	32.6 (324)
håp|håpet|håper|håpte|håpa	8.3 (109)	20.3 (266)	28.6 (375)	20 (63)	15.6 (49)	35.6 (112)

From the "Total" columns, it can be observed that HÅP(E), including all forms, is more frequently attested in match reports than in fiction in Norwegian (35.6 phtw vs. 28.6 phtw), whereas the opposite is the case for English HOPE, although only marginally so (33.5 phtw vs. 32.6 phtw). However, as is evident from Table 1, it is not merely a question of differences between the genres and languages, but also between word classes. This is visualised more clearly in Figure 4, where the marked differences in noun (green) vs. verb (blue) uses are fairly obvious. For the purpose of this visualisation percentages are used to illustrate the proportions of noun vs. verb uses. Although there are some outliers in the material, these do not significantly affect these proportions.[10]

[6] The ENPC+ corpus code identifies the author of the text (MoAl = Monica Ali), text number by that author (1) and language (E). The code of the Norwegian translation (T) of this text is MoAl1TN. See Ebeling and Ebeling (2013) for an overview of texts and text codes in the ENPC+.
[7] Developed by J. Ebeling and hosted by the University of Oslo (restricted access and password protected).
[8] The forms *håpt* (past participle), *håpende* (present participle) and *håpene* (definite noun, plural) were not attested in the ENPC+.
[9] An additional 11 instances of *Hope* as a proper noun have been left out of this study.
[10] For example, for verbs in English fiction there is one outlier. However, a Wilcoxon Rank Sum test (as implemented in R) shows that there is no significant difference in the material with and without this outlier (p=0.85); the same applies to the other sub-corpora that have between one and three outliers.

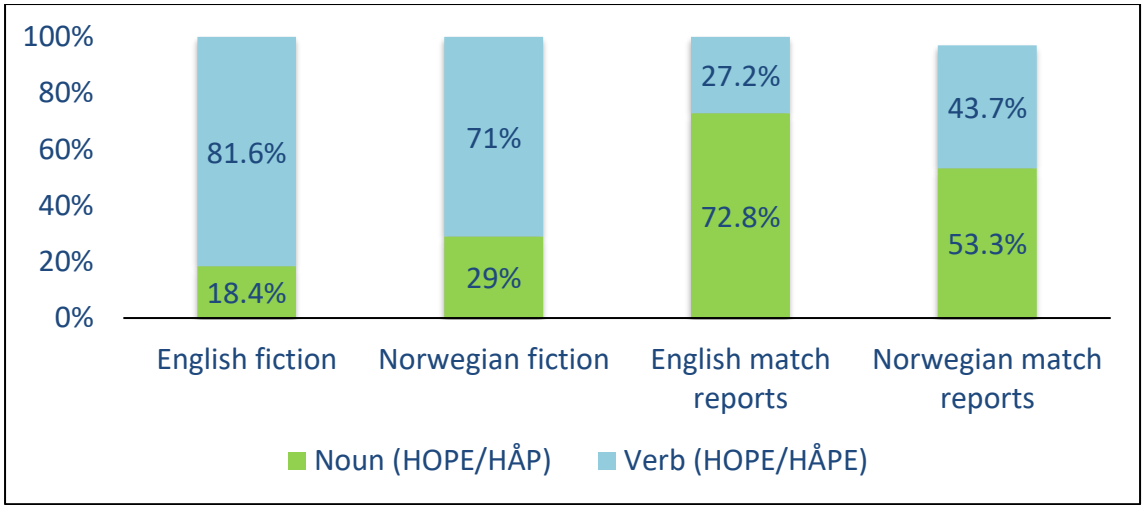

Figure 4. Distribution of noun and verb uses in the ENPC+ (fiction) and ENMaRC (match reports). (See Table 1 for raw numbers and normalised frequencies per 100,000 words.)

As can be observed in Figure 4, the distribution is more similar across the two languages than across the two genres, i.e. verbs are more common in fiction in both languages, whereas nouns are more common in match reports. These genre differences are in fact statistically significant for both nouns and verbs in English fiction vs. match reports and for nouns in Norwegian fiction and match reports.[11] This is in line with what Biber *et al.* (1999) note for English fiction and news, of which football match reports can be seen as a sub-category:

> The lexical word classes […] vary greatly across registers: Nouns are most common in news (and to a lesser extent in academic prose); they are by far the least common in conversation. […] Verbs and adverbs are most common in conversation and fiction. (Biber *et al.*, 1999: 65)

Figure 4 also suggests that these preferences (for verb in fiction and noun in match reports) are more prominent in English than in Norwegian. In other words, there is a narrower difference between the two word classes in the Norwegian material, particularly in the football match reports. This greater presence of verbs in the match reports may be related to what Nordrum (2007) notes in her dissertation on nominalizations in an English-Norwegian-Swedish contrastive perspective, namely that "there is a particularly strong and well-established prescriptive norm in Norway and Sweden favoring a 'verbal' or 'oral' style" (Nordrum, 2007: 219). This does, however, not explain the larger proportion of nouns in the Norwegian fiction material compared to English. And although the difference is not statistically significant for nouns in English vs. Norwegian fiction ($p=0.3117$), it is an observation that deserves further study in the future.

Following this general overview of noun and verb uses, the analysis now proceeds into the lexico-grammatical features of each word class in a comparison of their uses across the two languages and genres.

[11] Not all datasets were normally distributed, thus a Wilcoxon Rank sum test (in R) was chosen for the significance test, returning the following results: $p<0.0001$ for nouns and $p<0.001$ for verbs in English fiction vs. match reports, respectively, and $p<0.05$ for nouns in Norwegian fiction vs. match reports. The difference in verb uses in the Norwegian genres was not statistically significant ($p=0.1413$).

4. Cross-linguistic and cross-genre analysis of the noun and verb lemmas

Section 4.1 starts with an overview of the contextual features relevant to the English and Norwegian nouns in the material before moving on to a comparison of the features that stand out as being typical in each of the sub-corpora, i.e. English football match reports, Norwegian football match reports, English fiction and Norwegian fiction. Section 4.2 follows the same structure for the verbs.

4.1 The nouns HOPE and HÅP

To determine the phraseology of the nouns, the following contextual features were registered:

- **Form**:
    - Singular/Plural
- **Modification**:
    - Premodification (adjective | noun)
    - Postmodification (incl. apposition) (PP | *Ø-that* clause | *that*-clause | infinitive clause | relative clause)
- **Syntactic function**:
    - Head of NP and (part of) S | dO | sP
    - Head of NP and part of prepositional complement
- **Context** (negative | not negative)
- **Verb collocate**

Examples (5) and (6) serve to illustrate this classification scheme.

(5) I had brought with me a new *hope*. [ENPC+/BO1]

(6) ... og satte inn unggutten Erling Braut Håland i *håp* om å skape mer. I stedet var det ... [ENMaRC/VIF]
Lit: 'and brought on the young lad EBH in hope about to create more. Instead was it...'

In (5), *hope* is in the singular, premodified by the adjective *new* and head of a noun phrase functioning as the direct object. There is no evidence of a negative outcome, thus the context is deemed 'not negative', and the verb collocate, i.e. the verb in the clause containing *hope*, is *brought*. In the Norwegian example in (6), *håp* is also in the singular form, postmodified by a prepositional phrase (*om å skape mer* 'of creating more'; lit.: about to create more') and part of a prepositional complement following the preposition *i*. There are contextual clues suggesting that the context is negative (i.e. *Istedet var det* … 'instead it was'; lit.: instead was

it),[12] and there is arguably no verb collocate, as *skape* 'create' is part of a clause embedded within the postmodifying PP of *håp* and therefore not directly linked to it.

Table 2 gives a numerical overview of the selected contextual features in terms of proportions (i.e. percentages of total) within each sub-corpus, while Figure 5 visualises these according to the corpus model presented in Figure 3 (comparable parts).

Table 2. Main contextual features of the nouns HOPE and HÅP and their frequency (raw and percentage of total number of occurrences within each sub-corpus).

	English match reports		**Norwegian match reports**		**English fiction**		**Norwegian fiction**	
	Raw	% of total (235)	Raw	% of total (63)	Raw	% of total (83)	Raw	% of total (109)
PP postmodification	116	**49.4%**	30	**47.6%**	42	**50.6%**	54	**45.5%**
Premodification	40	17%	20	**31.7%**	15	18.1%	19	17.4%
Modification (pre or post)	155	**66%**	51	**81%**	57	**68.7%**	75	**68.8%**
HOPE/HÅP as part of S/dO/sP	181	**77%**	42	**66.7%**	59	**71.1%**	61	**56%**
Genre-specific verb collocates[13]	100 / 182	**55%**	28 / 42	**66.7%**				
General verb collocates[13]					40 / 56	**71.4%**	50 / 59	**84.7%**
Plural HOPE/HÅP	120	**51%**	0	0%	10	12%	4	3.7%
Negative contexts	116	**49.4%**	35	**55.6%**	40	**48.2%**	26	23.9%
HOPE/HÅP part of prep. complement	53	22.6%	19	**30.2%**	21	25.3%	42	**38.5%**

In Table 2, salient contextual features in the sub-corpora, represented as percentages, have been highlighted in bold and have been included in Figure 5. A feature is considered salient either if it is found in a minimum of ca. 50% of the cases, or if it is proportionally more frequent in a particular sub-corpus compared to the others, e.g. HÅP as part of a prepositional complement in the two Norwegian sub-corpora.

[12] It should be noted that, although it is difficult to objectively operationalise the feature of negative vs. non-negative context, the contextual clues are often quite clear in this regard.

[13] The percentages for verb collocates are calculated on the basis of a reduced number of occurrences, as verb collocates do not feature in instances where the nouns are part of a prepositional complement; cf. example (6). Thus, the total number of occurrences with verb collocates is reduced to 182 in the English match reports, to 42 in the Norwegian match reports, to 56 in the English fiction texts and to 59 in the Norwegian fiction texts.

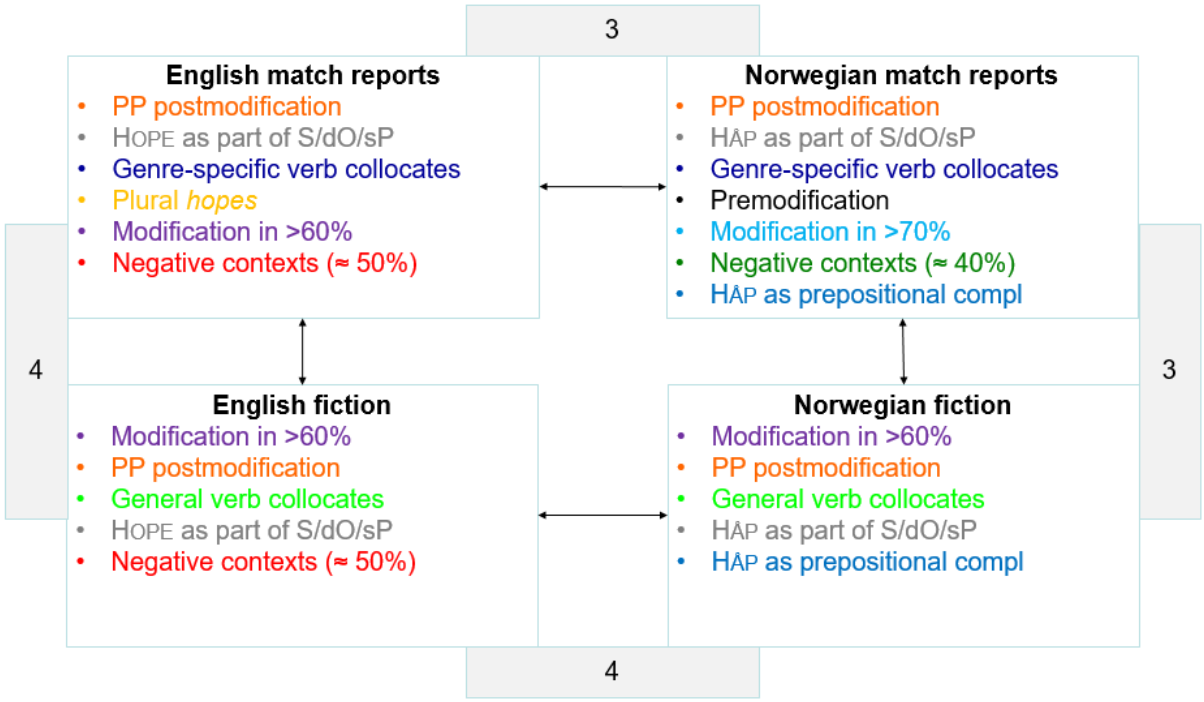

Figure 5. Contextual features of the nouns HOPE and HÅP: Main tendencies.

In Table 2 and Figure 5, we can observe that some characteristics are general for the two nouns in the two genres and languages:[14] postmodification by a PP, the noun is part of the S/dO/sP. A quantification of the similarities and differences is captured in the grey-shaded boxes connecting the different sub-corpora in Figure 5; these show the number of overlapping tendencies. There are three typical features that overlap between English and Norwegian football match reports (the two general features – PP postmodification and part of S/dO/sP – plus arguably more genre-specific verb collocates; see Figures 6 and 7) and three between Norwegian match reports and Norwegian fiction (the two general ones plus HÅP being part of a prepositional complement). There is even more similarity between the use of the nouns in fiction in the two languages, as well as between the genres in English, with four overlapping features each: modification in > 60% and general verb collocates in addition to PP postmodification and part of S/dO/sP for English and Norwegian fiction and modification in > 60%, and negative contexts ≈ 50% in addition to PP postmodification and part of S/dO/sP for English fiction and match reports. Thus, the features that set the genres or languages somewhat apart in the use of the nouns are type of verb collocates, degree of modification, proportion of negative contexts, the use of plural *hopes* and to some extent syntactic function (i.e. the Norwegian noun is more often found as part of a prepositional complement, typically in the sequence *i håp om* 'in hope about' ≈ 'in the hope that'). According to a Log-likelihood test,[15] the difference is statistically significant in the use of plural *hopes* between the two genres in English ($p<0.0001$ with a high effect size: Odds Ratio=16.25), in the use in negative contexts between the two genres in Norwegian ($p<0.0001$ with a small effect size: OR=0.18) and in the use of the noun as part of a prepositional complement between English and Norwegian fiction ($p<0.01$ with a small effect size: OR=2.05).

[14] In terms of dispersion, it should also be noted that most of the features – both for the nouns and verbs (see Table 3 and Figure 8) – are attested in most of the corpus files, albeit with a varying number of occurrences, particularly for features with a low number of attestations overall. A systematic look at dispersion would therefore be welcome in the future, preferably on a larger dataset.

[15] Using the log-likelihood calculator available at http://ucrel.lancs.ac.uk/llwizard.html.

Regarding type of verb collocate, the English and Norwegian match reports share the characteristic of making use of verb collocates that are arguably more (football-genre) specific in combination with HOPE (e.g. DASH HOPE, REIGNITE HOPE, QUASH HOPE), while the fiction texts share the feature of making use of more general verb collocates (e.g. HAVE HOPE, BRING HOPE). This difference in verb preferences between the genres becomes apparent in the word clouds in Figures 6 and 7 for English match reports and fiction, respectively.[16]

Figure 6. Genre-specific verb collocates (Eng. match reports). **Figure 7.** General verb collocates (Eng. fiction).

Figures 6 and 7 show some overlaps between the most frequent verb collocates in the two genres in English, but, not surprisingly, the verbs are generally different, and, as pointed out above, arguably more genre-specific and action-related in the football match reports. A very similar trend is noted for Norwegian, with verb collocates such as TENNE ≈ 'ignite/light', ØYNE ≈ 'see/nurture' and SVINNE ≈ 'vanish' in the match reports. Examples (7) and (8) serve to illustrate this cross-linguistic tendency of more genre-specific verbs in the match reports.

(7) ... but another defensive error *killed off* any *hope* of a comeback ... [ENMaRC/AFC]

(8) Scoringen *tente* et ørlite *håp*. [ENMaRC/STB]
Lit.: The goal lit a tiny hope

The fiction texts, on the other hand, tend to have more general verb collocates in both languages, with BE/VÆRE and HAVE/HA as the most prominent ones, e.g. examples (9) and (10).

(9) The poor devil didn't *have* a *hope* in hell. [ENPC+/PeRo1E]

(10) ... og det *er* vårt *håp* at de beste av våre landsmenn følger vårt eksempel. [ENPC+/BHH1]
Lit.: and it is our hope that the best of our countrymen will follow our example.

Returning to Figure 5 (and Table 2), we can further note that English HOPE (regardless of genre) occurs in negative contexts in roughly 50% of the cases, as evidenced in both examples (7) and (9), whereas HÅP is less often found in such contexts, particularly in Norwegian fiction with roughly 23% of the cases; neither example (8) nor (10) is deemed negative.

4.2 The verbs HOPE and HÅPE

Following the procedure applied to the nouns in section 4.1, the following contextual features were recorded for the verbs:

[16] The word clouds were generated in WordArt.com on the basis of a list of all verb collocates occurring more than once in the corpora; see the Appendix for the number of actual occurrences in each case and that determine the size of the verbs in the clouds.

- **Verb form**:
    - Tense, aspect, modality, voice
- **Verb complementation**:
    - *Ø-that/at* clause | *that/at*-clause | infinitive clause | PP (*for/på*) | intransitive use (no complementation) | prop word | parenthetical use (…, *I hope* | …, *håper jeg*)
- **Subject**:
    - Pronoun | NP (including proper nouns)
- **Context** (negative | not negative)

The classification framework is illustrated in examples (11) and (12), where the former is a relatively typical example of English fiction: the personal pronoun *I* is the subject, *hope* is in the present tense, it is followed by a *Ø-that* clause and the context is not negative. The translation into Norwegian is included in example (11) and demonstrates a highly congruent – almost word for word – rendering, to illustrate that this is also typical of Norwegian fiction. Similarly, example (12) is a relatively typical example of the English football match reports, with a full NP as subject, *hoping* is the main verb in a modal perfect progressive verb phrase followed by a *Ø-that* clause and the context is negative.[17]

(11) I *hope* everything goes well for you. [ENPC+/AnCl1E]
Jeg *håper* alt går bra for deg. [ENPC+/AnCl1TN]

(12) The Head Coach *would have been hoping* his team could hold out until half-time... [ENMaRC/WFC]

An overview of the distribution of the contextual features in each sub-corpus is given in Table 3, while Figure 8 visually summaries the main tendencies.

Table 3. Main contextual features of the verbs HOPE and HÅPE and their frequency (raw and percentage of total number of occurrences within each sub-corpus).

	English match reports		**Norwegian match reports**		**English fiction**		**Norwegian fiction**	
	Raw	% of total (88)	Raw	% of total (49)	Raw	% of total (367)	Raw	% of total (266)
Simple present tense	3	3.4%	28	**57.1%**	136	**37.1%**	126	**47.4%**
Simple past tense	4	4.5%	7	14.3%	90	**24.5%**	87	**32.7%**
Past perfect	7	8%	8	**16.3%**	19	5.2%	22	8.3%
Non-finite (inf./*-ing*) clause	32	**36.4%**	6	12.2%	84	22.9%	30	11.3%
Progressive aspect	34	**38.6%**	N/A	-	39	10.6%	N/A	-
Ø-that/at clauses	25	28.4%	20	**40.8%**	216	**58.9%**	106	**39.8%**
That/at-clauses	4	4.5%	11	**22.4%**	30	8.2%	87	**32.7%**

[17] It is interesting to note that HOPE occurs in the progressive aspect much more frequently in the match reports than in the fiction texts, in 34 out of 88 cases (38.6%) vs. 39 of 367 (10.6%), respectively. Several scholars have pointed to an increased use of the progressive with stative verbs or in combination with modal verbs in recent years (Aarts *et al*., 2010; Leech *et al.*, 2009). However, it is hard to determine, and also beyond the scope of this study, whether there is a genre or diachronic effect here.

Inf.-clause complementation	31	**35.2%**	6	12.2%	49	13.4%	10	3.8%
Prep.complementation	24	**27.3%**	11	**22.4%**	32	8.7%	24	9%
NP Subject	58	**65.9%**	3	6.1%	36	9.8%	24	9%
1st p pl Subject	14	15.9%	23	**46.9%**	20	5.4%	18	6.8%
1st and 3rd p sg Subject	5	5.7%	16	32.6%	277	**75.5%**	193	**72.6%**
Negative contexts	18	**20.5%**	11	**22.4%**	55	15%	44	16.5%

As was the case in Table 2 for the nouns, the most salient contextual features in the sub-corpora (percentages) have also been highlighted in bold in Table 3 for the verbs and are included in Figure 8.

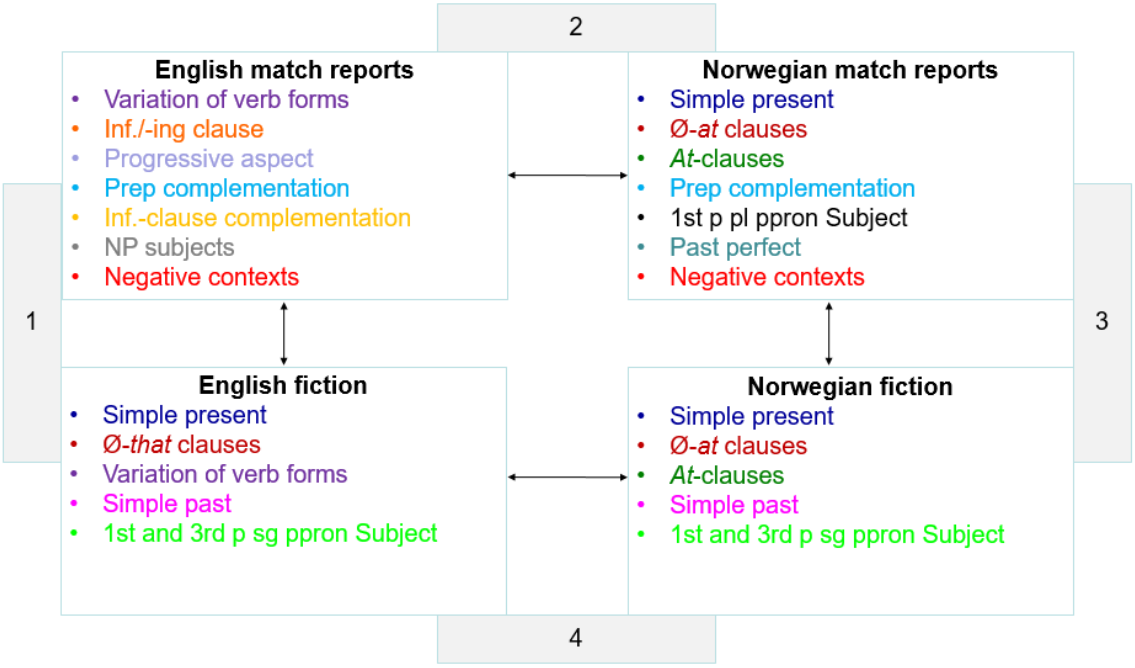

Figure 8. Contextual features of the verbs HOPE and HÅPE: Main tendencies.

Figure 8 shows that there is less overlap of typical features between the sub-corpora here than was the case for the nouns (see Figure 5), notably with only one overlapping feature for English fiction and match reports and only two for English and Norwegian match reports. There is most similarity between the fiction texts in English and Norwegian, which are characterised by Ø-*that*/*at* clauses, a combination of the simple present and past tense, and 1st and 3rd person singular pronouns as Subjects. The Norwegian texts also show a fair amount of overlap across the two genres, whereas there seems to be more of a text type effect in English. There are few features that are typical of both English match reports and English fiction.[18] In fact, they only share one of the characteristic features that can be gleaned from Table 3 for the verb HOPE, namely variation of verb forms. By variation is here understood a combination of tense, modality and aspect, as well as the use of non-finite forms. A couple of examples are given in (13)–(15), with a modal perfect, a present perfect progressive and a non-finite *-ing*, respectively.

[18] Statistically significant differences were recorded for the following features, according to a LL test: simple present tense ($p<0.0001$; OR=0.03), simple past tense ($p<0.0001$; OR=0.06), non-finite clauses ($p<0.0001$; OR=0.52), Ø-*that* clauses ($p<0.0001$; OR=0.16) and negative contexts ($p<0.01$; OR=0.44). However, it is important to note that, in some cases, these tests are based on very small numbers.

(13) ... and was unable to execute his volley in the way he *would have hoped* ... [ENMaRC/CFC]

(14) *Have* you *been hoping* for something more exciting? [ENPC+/ABR1]

(15) Bradley sent his team out for the restart *hoping* they could find a way of causing more problems for the Watford defence. [ENMaRC/SCAFC]

Figure 8 also reveals a low degree of overlap between English and Norwegian match reports,[19] as they only share two typical contextual features for HOPE and HÅPE, namely prepositional complement and negative context, both of which are captured in example (16).

(16) Det ble ikke den festkvelden vi hadde *håpet på* Aker Stadion søndag kveld. [ENMaRC/MFK]
Lit.: It did not become the night of celebration we had hoped for at Aker Stadion Sunday night

With reference to example (16) it is interesting to note that while genre seems to be a decisive factor for the verb to (more typically) be used in negative contexts, language was a decisive factor for the nouns, where the English match reports and English fiction texts were seen to be more in agreement regarding this feature.

5. Discussion

On the basis of preliminary observations of the data it was suggested in Section 1 that the match reports would make use of a narrower selection of contexts in which the 'hope' lemmas are used (cf. Stubbs 1996: 89). The case studies presented in Sections 4.1 and 4.2 do not seem to substantiate this hypothesis. In fact, from the potential uses attested in the corpora, the match reports feature a broader repertoire of typical phraseological contexts compared to fiction. Thus, instead of featuring in a narrower selection of contexts from the pool of potential uses, the lemmas are rather shown to typically feature in a different selection of contexts in the match reports. In terms of number of characteristic phraseologies recorded for both the nouns and verbs, genre seems to play a slightly more important role than language.

It could be argued that the contextual features recorded for the nouns and verbs are relatively straightforward to determine, perhaps with one exception: 'context'. A binary distinction of negative|not negative was applied to extended contexts in which HOPE and HÅP(E) occurred. One potential challenge, also referred to above, was how to operationalise this in the analysis of individual instances, as HOPE and HÅP(E) are arguably reserved for positive connotations, indicating that the negative flavour with which these items are sometimes imbued seems almost contradictory. Example (17) is a case in point, where *surge of hope* carries positive expectations that are later shown not to be fulfilled, when it turns out that it was not Emma who called, but Andrew. It may be speculated that this is a deliberate choice on the part of the writer to create an effect, i.e. an element of surprise, or as Louw (1993: 30) puts it "irony in the text or insincerity in the writer", with reference to the concept of semantic prosody (see further below).

[19] Syntactic differences between the two languages can be seen to account for some of this, as Norwegian does not have forms corresponding to non-finite *-ing* clauses and a grammaticalised progressive aspect. Statistically significant differences can be noted for the following features: simple present tense ($p<0.0001$;OR= 0.03), past perfect ($p<0.05$; OR=0.28), *Ø-at/that* clause ($p<0.01$; OR=0.40), *at/that*-clause ($p<0.0001$; OR=0.12), NP Subjects ($p<0.0001$; OR=6.13) and 1st p. pl subject ($p<0.0001$; OR=0.19). However, it is important to stress that in some cases the number of occurrences are few here and we should perhaps not put too much weight on these tests.

(17) He had a *surge of hope* that it was Emma, until he picked it up and heard Andrew babbling excitedly... [ENPC+/MiWa1E]

Sometimes these contextual clues lie outside the scope covered by the default length of a concordance line, and a wider context has to be examined. It is also tempting to suggest that the past tense can be seen as a trigger for hopes being shattered. However, the evidence for this is inconclusive, as the past tense is also regularly used in non-negative contexts – either neutral or positive, as in (18) – where there is nothing in the surrounding context to suggest that the existence of hope came to an end.

(18) There was *hope* everywhere. [ENPC+/JSM1]

The contextual features recorded for the nouns and verbs bear a strong resemblance to features covered by the categories that are part of Sinclair's (1996, 1998) Extended Units of Meaning model, viz. collocation, colligation, semantic preference and semantic prosody. And although the material at hand does not uncover strong unanimous lexico-grammatical patterns for neither the nouns nor the verbs that would suggest that HOPE and HÅP(E) clearly function as cores of extended units of meaning, it does reveal tendencies regarding semantic prosody, which is the only obligatory element in the model apart from the core (Sinclair, 1998: 20; Ebeling and Ebeling, 2013: 58). Traditionally semantic prosody refers to semantic colouring through surrounding context and may contribute to a positive or negative reading of words that are in themselves neutral. Louw (1993: 157) defines semantic prosody as "a consistent aura of meaning with which a form is imbued by its collocates". Put differently, and according to Louw and Milojkovic (2016) collocates contribute to a "context of situation revealing attitude (semantic prosody)" (Louw and Milojkovic, 2016: 54). In the literature, it has been argued that such an attitude may be binary, i.e. positive vs. negative, or non-binary (and more specific), expressing e.g. 'difficulty' in the case of *the naked eye* (Sinclair 1996: 33) and 'occupation' in case of *train as a* (Hoey, 1997: 5). For the purpose of this study, the binary opposition negative vs. non-negative has been applied.[20] In the context of the current investigation it is also important to mention that several scholars have observed that semantic prosody may be both language-specific (e.g. Stewart, 2009: 32)[21] and register-specific (Xiao and McEnery, 2006: 114ff; Hunston, 2007: 263ff).

For the items under investigation here, there seems to be (at least) two things at play regarding semantic prosody: language and genre. In the case of the nouns, the strongest indication of a negative-like prosody is found in English, regardless of genre (see Table 2 and Figure 5). For the verbs, on the other hand, it is the match reports that show the strongest tendency towards a negative prosody, regardless of language (see Table 3 and Figure 7). The verbs in fiction do not seem to take on a particular prosody at all, as most instances seem to contain a neutral use of HOPE and HÅPE, as in examples (19) and (20).

(19) Though what you *hope* to find there I have no idea. [ENPC+/PeRo2E]

(20) Fortsatt *håpet* jeg på Kari Thue. [ENPC+/AnHo2N]
Lit.: Still I hoped for Kari Thue

The Norwegian noun behaves differently from the English noun in being less consistently used in negative contexts. However, it is clear that HÅP is closer to an established negative prosody in the match reports (occurring in negative contexts in ca. 40% of the cases), e.g. example (21),

[20] It is important to note that the validity of semantic prosody as a concept has been questioned over the years, but it would take us too far afield to go into this discussion here. But see e.g. Whitsitt (2005), Hunston (2007), Morley and Partington (2009) and Stewart (2010) for some (critical) discussions.

[21] In the case of mismatched prosodies across languages, see also Partington (1998), Tognini-Bonelli (2002) and Ebeling (2014), and references therein.

than it is in the fiction texts, where it occurs in negative contexts in roughly 23% of the cases (see Table 3); this difference was shown to be statistically significant, albeit with a small effect size.

(21) Alt *håp* om poeng ser nå ut til å være over. [ENMaRC/OBK]
Lit.: All hope of points now looks to be over

In a few instances HÅP is used in contexts with a positive outcome, as example (22) arguably illustrates – a hope has been restored after it had been dashed –, but it is by far most commonly used in more neutral contexts, expressing a hope with expectations of a positive outcome, but where the outcome is in fact unknown, as in (23).

(22) Likevel hadde jeg det bedre en stund. Både fordi Henrik hadde gitt meg tilbake et *håp* jeg ikke lenger hadde … [ENPC+/MN1]
Still I felt better for a while. Both because Henrik had given me back a hope I no longer had … [ENPC+/MN1T]

(23) For det var det eneste svaret som ga noe *håp*. [ENPC+/JoNe1N]
Because it was the only answer that gave any hope. [ENPC+/JoNe1TE]

Even in contexts in which the immediate collocates are of a positive nature, as in (24) where the positive adjective *godt* 'good' premodifies *håp*, the outcome is not specified as positive in the surrounding context.

(24) "Jeg tror jeg verken skal bekrefte eller avkrefte annet enn at vi i Kripos har godt *håp* om at denne saken går mot en snarlig oppklaring." [ENPC+/JoNe2N]
"I don't think I have to confirm or deny anything except that we at Kripos are fairly confident [Lit.: … Kripos have a good hope ...] that we will soon have this case solved." [ENPC+/JoNe2TE]

Returning to the starting point of this study, and to the question of whether the prominent use of HOPE/HÅP(E) in negative contexts in match reports (Ebeling, 2019) reflects a true tendency of this genre in both languages and whether such use extends to other genres, we can conclude that the investigation uncovers some conflicting evidence in this respect. Both genre and language seem to have an impact, thus lending some support to the observations referred to above, namely that semantic prosody may be both language- and register-specific. However, it is also interesting to note that not only may semantic prosody depend on language and genre/register, it may also be dependent on word class. This is in accordance with previous studies that have shown similar trends, e.g. Stubbs (1995) in the case of the noun and verb CAUSE.[22]

In summary, then, the attraction to negative contexts seems to be language-specific for the nouns HOPE/HÅP, whereas it seems to be genre-specific for the verbs HOPE/HÅPE.

6. Concluding remarks and suggestions for further study

In addition to investigating the potentially negative bias of HOPE/HÅP(E), this study set out to answer a set of research questions regarding the use of these cognates in two languages and two genres on the basis of two different kinds of contrastive corpora. To answer the specifically cross-linguistic question – to what extent are the lemmas used similarly in English and

[22] The slight difference noted in the semantic prosody between the verb and noun CAUSE is very much tied to one specific inherent meaning of the noun, namely the 'aim/principle' reading, as in *The only* cause *they had in common was a refusal to eat meat*. (ENPC/PDJ3). The importance of taking separate meanings into account when investigating semantic prosody has been addressed in a recent master's thesis by Russnes (2020).

Norwegian? – the investigation reveals that, in very general terms, the lemmas show similar potential of use but have slightly different preferred uses. An almost identical answer is suggested for the specifically cross-genre question – to what extent are the lemmas used similarly in match reports and fiction? In this case the lemmas have a relatively similar potential of use, but with different preferences in the two genres studied.

In more specific terms, the nouns are shown to be more similar across the genres in English with more overlapping contextual features than in Norwegian. In fact, Norwegian fiction is more similar to English fiction than to Norwegian match reports (see Table 2 and Figure 5). The analysis of salient contextual features further suggests that the Norwegian match reports seem to have adopted a slightly different set of salient features compared to the other sub-corpora. As far as the verbs are concerned, their use is more similar across the languages than across the genres, (particularly in fiction). Moreover, Norwegian fiction and match reports are more similar than English fiction and match reports (see Table 3 and Figure 8). Thus, in this case, it is the English match reports that seem to have adopted the most special uses. A final and general observation from the case studies is that, overall, the use of these lemmas is most similar in English and Norwegian fiction. At this stage we may only speculate as to the reason for this, but could it be that we are dealing with an established (fiction) genre versus an emerging and less established genre of online match reports, which, as a result, produces more variation?

Regarding the third research question – to what extent does the use of different types of contrastive corpora contribute to our cross-linguistic knowledge of the lemmas? – it has been shown that the combination of comparable data in two genres and translation data in one genre has:

- Provided a firm basis for a contrastive analysis of the items compared;
- Highlighted language similarities/differences/preferences;
- Highlighted cross-linguistic genre similarities/differences/preferences.

These points may form the basis of a slightly modified version of Aijmer and Altenberg's (1996) frequently quoted words on the usefulness of parallel corpora, thus:

> The use of different contrastive corpora has given new insights into the languages and genres compared – insights that would have gone unnoticed in a study of only one of the corpora alone, i.e. either the ENPC+ or the ENMaRC.[23]

The advantages of drawing on several primary contrastive sources are evident, and in particular a combination of both bidirectional translation data and monolingual comparable data between two or more languages (representing different genres) has been shown to be fruitful. Such corpora have previously been shown to complement each other in the sense that bidirectional translation corpora (i.e. parallel corpora) arguably provide the researcher with a more objective *tertium comparationis*, while comparable corpora are indispensable in a contrastive study in order to provide data sets that are both more extensive and that include text types that are not typically translated between languages (cf. Johansson, 2007; Ebeling and Ebeling, 2020).

Against this backdrop, it is important to stress that the potential of neither corpus type has been exploited to the full in the current paper. Further insights could be gained on the basis

[23] The original Aijmer and Altenberg quote reads as follows: bilingual corpora "give new insights into the languages compared – insights that are likely to be unnoticed in studies of monolingual corpora" (Aijmer and Altenberg, 1996: 12).

of the ENPC+ in a systematic investigation of correspondences of noun and verb patterns in a contrastive perspective (only fiction), for example:

- Analyse the translation paradigms of the English verb pattern *hope* + Ø-*that* clauses, on the basis of examples such as (25);
    - Ø-*that* clause → *at*-clause, even if Ø-*at* clauses are possible in Norwegian

(25) He *hopes* [Ø] they'll have enough warm clothes to last the coming winter. [ENPC+/StGa1E]
Han *håper at* de har nok varmt tøy til å klare seg gjennom vinteren som står for døren. [ENPC+/StGa1TN]
Lit.: He hopes that they …

- Analyse the translation paradigms of the Norwegian noun pattern PREP *håp* PP + *at* clause, on the basis of examples such as (26).
    - PREP *håp* PP + *at* clause → non-finite *-ing* clause + Ø-*that* clause, where a syntactically similar pattern is ruled out in English

(26) ... *i håp om at* den andre skulle si navnet sitt; [ENPC+/EFH1]
Lit.: in hope about that ...
… *hoping* [Ø] the other man would say his name. [ENPC+/EFH1T]

Furthermore, the advantage of size that is often attributed to comparable monolingual corpora when compared to parallel corpora is not present here, with the ENMaRC being smaller than the ENPC+. However, this may be amended in the future, as more match reports can be added and more sizeable comparable fiction data can be culled from larger monolingual corpora in the two languages.

Finally, it should be noted that there are some challenges involved when trying to carry out a comprehensive case study on the basis of material from a variety of sources. In particular, it is challenging to organise and analyse data from multiple languages and genres in a clear and consistent manner. Nevertheless, rather than shy away from the relatively complex nature of such data, researchers should perhaps complement traditional contrastive analysis techniques with more sophisticated models for handling complex data, as it would give us the opportunity to gain even more rewarding insights into cross-linguistic, cross-genre uses of language.

References

Aarts, B., Close, J. and Wallis, S. 2010. Recent Changes in the Use of the Progressive Construction in English. In *Distinctions in English Grammar*, B. Capelle and N. Wada (eds), 148–168. Kaitakusha: Tokyo, Japan.

Aijmer, K. and Altenberg, B. 1996. Introduction. In *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*. *Lund 4–5 March 1994*, K. Aijmer, B. Altenberg and M. Johansson (eds), 11–16. Lund: Lund University Press.

Altenberg, B. 1999. Adverbial Connectors in English and Swedish: Semantic and Lexical Correspondences. In *Out of Corpora. Studies in Honour of Stig Johansson*, H. Hasselgård and S. Oksefjell (eds), 249–268. Amsterdam: Rodopi.

Anthony, L. 2019. AntConc (version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software [Last accessed 5 May 2020].

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

Chesterman, A. 1998. *Contrastive Functional Analysis*. Amsterdam: Benjamins.

Det Norske Akademis ordbok (Norwegian Academy Dictionary). 2020. Oslo: Det Norske Akademi for Språk og Litteratur. https://www.naob.no/ordbok [Last accessed 17 August 2020].

Dupont, M. and Zufferey, S. 2017. Methodological Issues in the Use of Directional Parallel Corpora. *International Journal of Corpus* Linguistics 22(2), 270–297.

Ebeling, J. and Ebeling, S.O. 2013. *Patterns in Contrast*. Amsterdam: Benjamins.

Ebeling, S.O. 2014. Cross-linguistic Semantic Prosody: The Case of *Commit*, *Signs of* and *Utterly* and their Norwegian Correspondences. In *Corpus-based Studies in Contrastive Linguistics*, *Oslo Studies in Language* 6(1), 2014, S.O. Ebeling, A. Grønn, K.R. Hauge and D. Santos (eds), 161–179.

Ebeling, S.O. 2019. The Language of Football Match Reports in a Contrastive Perspective. In *Corpus Approaches to the Language of Sports. Text, Media, Modalities*, M. Callies and M. Levin (eds), 37–62. London: Bloomsbury Academic.

Ebeling, S.O. 2021. Minutes of Action! A Contrastive Analysis of Time Expressions in English and Norwegian Football match reports. To appear in *Time in Languages, Languages in Time*, A. Čermáková, T. Egan, H. Hasselgård and S. Rørvik (eds). Amsterdam: Benjamins.

Ebeling, S.O. Forthcoming. The Function of Recurrent Word-combinations in English Translations from Three Different Languages. To appear in a special issue of *Meta: Translators' Journal*, ed. By C. Ji and M. Oakey.

Ebeling, S.O. and Ebeling, J. 2017. A Functional Comparison of Recurrent Word Combinations in English Original vs. Translated Texts." *ICAME Journal 41*, 31–52.

Ebeling, S.O. and Ebeling, J. 2020. Contrastive analysis, *tertium comparationis* and corpora." *Nordic Journal of English Studies* 19(1), 97–117.

Hoey, M. 1997. From Concordance to Text Structures: New Uses for Computer Corpora. In *Practical Applications in Language Corpora*, B. Lewandowska-Tomaszyk and P.J. Melia (eds), 2–23. Łódź: Łódź University Press.

Hunston, S. 2007. Semantic Prosody Revisited. *International Journal of Corpus Linguistics* 12(2), 249–268.

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

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

Leech, G., Hundt, M., Mair, C. and Smith, N. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: CUP.

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

Louw, B. 1993. Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies." In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis and E. Tognini-Bonelli (eds.), 157–175. Amsterdam: Benjamins.

Louw, B. and Milojkovic, M. 2016. *Corpus Stylistics as Contextual Prosodic Theory and Subtext*. Amsterdam: John Benjamins.

Morley, J. and Partington, A. 2009. A few Frequently Asked Questions about Semantic – or Evaluative – Prosody." *International Journal of Corpus Linguistics* 14(2). 139–158.

Neumann, S. 2014. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: De Gruyter Mouton.

Nordrum, L. 2007. *English Lexical Nominalizations in a Norwegian-Swedish Contrastive Perspective*. Doctoral dissertation, English department, Göteborg University. Available at https://gupea.ub.gu.se/bitstream/2077/17181/5/gupea_2077_17181_5.pdf [Last accessed 17 August 2020].

Partington, A. 1998. *Patterns and Meaning. Using Corpora for English Language Research and Teaching*. Amsterdam: Benjamins.

Oxford English Dictionary (OED) Online. 2020. Oxford: Oxford University Press <http://oed.com/> [Last accessed 17 August 2020].

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ [Last accessed 17 August 2020].

Russnes, M.S. 2020. *Committed to Crime – A Corpus-based Study of the Semantic Prosodies of Separate Meanings within Lexical Items*. Unpublished MA thesis, University of Oslo.

Sinclair, J. 1996. The Search for Units of Meaning. *Textus IX*. 75–106.

Sinclair, J. 1998. The Lexical Item. In *Contrastive Lexical Semantics*, E. Weigand (ed.), 1–24. Amsterdam: John Benjamins.

Stewart, D. 2010. *Semantic Prosody. A Critical Evaluation*. New York / London: Routledge.

Stubbs, M. 1995. Collocations and Semantic Profiles. On the Cause of Trouble with Quantitative Studies. *Functions of Language*, 2:1, 23–55.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

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

Tognini-Bonelli, E. 2002. Functionally Complete Units of Meaning across English and Italian: Towards a Corpus-driven Approach." In *Lexis in Contrast. Corpus-based Approaches*, B. Altenberg and S. Granger (eds), 73–95. Amsterdam: John Benjamins.

Whitsitt, S. 2005. A Critique of the Concept of Semantic Prosody. *International Journal of Corpus Linguistics* 10/3. 283–305.

Xiao, R. and McEnery, T. 2006. Collocation, Semantic Prosody, and Near Synonymy. A Cross-linguistic Perspective. *Applied Linguistics* 27 (1): 103–129.

Appendix

Table A. Number of occurrences of verb collocates with a frequency of more than one in the English match reports and fiction texts.

English match reports	
Verb	**Raw freq.**
BE	10
BOOST	4
DASH	17
DEAL	3
END	4
EXTINGUISH	6
GIVE	43
GO	5
HANG	2
HAVE	2
HIT	2
KEEP	15
KILL OFF	8
OFFER	5
PROVIDE	2
QUASH	2
RAISE	5
REIGNITE	3
SUFFER	4
TAKE	3
THWART	2

English fiction	
Verb	**Raw freq.**
BE	12
BRING	2
GIVE	3
GIVE UP	4
HAVE	10
HOLD OUT	5
MAKE	2
RETURN	2

Author's address

Signe Oksefjell Ebeling
Department of Literature, Area Studies and European Languages
University of Oslo
P.O. box 1003, Blindern
NO-0315 Oslo
Norway
s.o.ebeling@ilos.uio.no

Prepositional phraseological patterns in Czech and English

Towards a contrastive study resource[1]

Denisa Šebestová

Charles University, Prague (Czech Republic)

This pilot study aims to identify differences in native and non-native phraseologies, focussing on prepositional patterns. Previous research suggests L2 users' limited phraseological choices may hinder the accuracy of their language production, and prepositions can pose a particular challenge to Czech learners of English, given the lack of correspondence between translation equivalents. Further, prepositional patterns contribute to text structuring, making them an important part of learners´ competence. Using representative corpora of English and Czech, 3- to 5-grams containing the equivalent preposition pair *in*/*v* are extracted. The identified patterns are classified by their semantics and textual functions. While *in/v* patterns mostly fulfil corresponding functions in the languages compared, the distribution of these functions differs. Specifically, some pattern types are only found in English, highlighting its analytic nature as opposed to inflectional Czech.

Keywords: n-grams, prepositions, native and non-native phraseology, typologically distant language pair, Czech/English

1. Introduction

This study is based in cross-linguistic distributional (Granger and Paquot, 2008) or data-driven (Granger and Meunier (eds), 2008) phraseology, i.e. examining recurrent word combinations through corpora. It was prompted by earlier findings provided by research into non-native phraseology (Ebeling and Hasselgård, 2015; Granger, 2017; Granger and Bestgen, 2014; Hasselgård, 2019; Vašků, Brůhová, and Šebestová, 2019), as well as by the interest in – and need for – teaching materials reflecting those findings (Reppen, 2011). It is conceived as a pilot study, aiming to contrast a selected pattern group – prepositional patterns – between the typologically distant language pair of Czech and English. The results of this contrastive analysis can then be used as a springboard towards suggesting how n-gram based studies of phraseology can inform foreign language instruction.

[1] This research was funded by the Faculty of Arts, Charles University, within the project 'Specifický vysokoškolský výzkum - Jazyk a nástroje pro jeho zkoumání' (2020).

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.v11i1.3437.

Section 2 introduces the theoretical background and motivation for the study. Section 3 introduces the material and methods employed in the study. Section 4 presents the textual functions conveyed by prepositional patterns in the English and Czech data. Results are described for each language separately. Section 5 reports on differences in pattern usage between the two languages. Finally, Section 6 summarizes the results and suggests potential avenues for further research.

2. Background and motivation

Phraseology (in the sense of the use of recurrent word combinations, cf. Gray and Biber, 2015: 125; Ebeling and Hasselgård, 2015: 207) has been shown to "unmistakably [distinguish] native speakers of a language from L2 learners" including advanced learners (Granger and Bestgen, 2014:,229). It has been suggested that L2 learners have a limited repertoire of phraseological sequences, and employ these in ways which differ considerably from native usage (Granger, 2017). As a result, L2 learners tend to overuse a restricted set of phraseological sequences which they have mastered and feel confident using. Hasselgård (2019) terms these 'phraseological teddy bears', referring back to Hasselgren's (1994) idea of 'lexical teddy bears'.

These limitations have a serious bearing on the learner's communicative skills: they pose a potential hindrance to language production, since phraseological competence forms a crucial part of a learner's overall language proficiency (Howarth, 1998; Hyland, 2008; Paquot, 2018; Paquot and Granger, 2012). The degree of phraseological competence is also an important criterion in determining L2 fluency, distinguishing native speakers from non-native learners (Granger and Bestgen, 2014; Hasselgård, 2019). Moreover, becoming acquainted with recurrent word combinations is important as they form a major component of everyday language use (Biber *et al*., 2004; Erman and Warren, 2000).

One way to address this issue is to contrast the phraseologies of the target and source languages, using the results to inform language instruction. For instance, Granger (2018) combines contrastive analysis (comparing different languages) with a translation studies perspective and learner corpus data. The resulting 'Contrastive Translation Analysis' approach allows for comparing original language to translated, as well as learner language to native, and by extension "to tease out developmental vs. L1-specific features of interlanguage" (Granger, 2018: 4). This suggests that a contrastive corpus analysis can produce valuable insights into how a speaker's knowledge of their L1 can be reflected in their L2 production. Granger also points out the value of phraseology for examining the influence of one language on another, including L1 transfer in learner language (ibid.). She concludes that frequent phraseological combinations, which can be efficiently unveiled through n-gram extraction, are of great relevance to L2 learners (ibid.: 5), in line with studies of phraseological competence (Paquot, 2018 among others).

Aiming to contribute to the contrastive description of phraseology, the present study compares the use of patterns containing the equivalent preposition pair *in – v* between English and Czech. The results should ultimately inform a study resource developing the phraseological competence in advanced Czech students of English, primarily at university level. Further, the phraseological contrastive analysis of this language pair is potentially valuable from the typological perspective. Previous cross-linguistic phraseological studies indicate that the n-gram method can efficiently identify recurrent sequences and point out cross-linguistic differences in their use. However, n-grams pose methodological difficulties when dealing with typologically distant language pairs, such as English and Spanish, French, or Czech (Čermáková and Chlumská, 2016, 2017; Cortes, 2008; Granger, 2014; Šebestová and Malá,

2019). In the case of Czech, the challenges are due to the highly inflectional nature of Czech, as opposed to predominantly analytical English. A further obstacle is posed by the greater variability of Czech word-order compared to English. Both these factors influence the delimitation of a recurrent multi-word unit in Czech, and have motivated the development of new software capable of identifying patterns with partial lemmatisation and positional mobility (cf. Section 3).

2.1 Prepositional patterns

As pointed out by Hunston (2008), focus on phraseological patterns containing grammatical words ('small words', ibid.) can be beneficial because such patterns contribute to shaping the structure of texts. Fulfilling important textual functions, on a larger scale these grammatical patterns also help reveal pervasive discourse patterning. Discourse-organizing functions are frequently fulfilled by phraseological combinations (Granger, 2018:6), which further indicates that the n-gram method is a suitable means to this end. Moreover, discourse organizing and text structuring is a crucial skill for advanced learners (Granger, 2018), especially for university students, required to produce complex written assignments. Hence, a 'small words' approach seems suitable for this study. Another argument in favour of using grammatical words as the starting point is their extensive frequency and dispersion throughout discourse (Groom, 2010; Sinclair, 1991), making them an efficient tool to provide a comprehensive portrait of the phraseological characteristics of a corpus, to identify a variety of pattern types fulfilling different textual functions and manifesting varying degrees of formulaicity (Groom, 2010:71). For these reasons, function words seem a valid starting point for this study.

Specifically, prepositions were selected as the basis for the identification of phraseological patterns. Prepositions are a valuable starting point from the contrastive and pedagogical perspective since they are a frequent source of errors in EFL students, including advanced learners; apart from their polysemy and polyfunctionality, this is possibly due to a large degree of translation non-correspondences, and inaccurate/oversimplified representation in translation dictionaries (Klégr and Malá, 2009; Peřestá, 2017). In this pilot study, I focus on the preposition pair *in – v*, ranking among the most frequent prepositions in both languages. To summarize, this study aims to identify prepositional patterns involving the translation equivalent preposition pair *in – v* in representative corpora of English and Czech, respectively. These patterns will be described in terms of their textual functions and compared across the two languages.

Although *in* and *v* are translation equivalents, their senses and contexts of use do not entirely correspond across languages (Klégr and Malá, 2009; Peřestá, 2017). The polysemic nature of prepositions seems an important factor, as different senses of a preposition will often be translated by different equivalents (Klégr and Malá, 2009). Consequently, both *in* and *v* are likely to fulfil a range of textual functions. However, the functions carried out by each preposition are expected to differ between the two languages. My aim is to inquire into the nature and extent of these cross-linguistic differences.

2.2 Corpus methods in language teaching

Interest in corpus-informed teaching materials has been growing and influencing approaches to foreign language instruction (Huang, 2011; Reppen, 2011). Corpus material can help learners become acquainted with authentic language, presenting them with a variety of contexts of use (Reppen, 2011:35). Reppen outlines three techniques of employing corpora in language instruction: learning aids prepared by the instructor based on corpus data; interactive practice with students using corpora in class; and using (available or custom-made) specialized corpora

(2011: 36), enabling learners "to explore the patterns found in the writing of their discipline" (2011: 44). Likewise, Hyland (2008: 5) points out the importance of advanced learners knowing discipline-specific phraseological expressions, since "their very 'naturalness' [signals] competent participation in a given community". His analysis shows that scientific disciplines are distinguished by their use of patterns. These patterns are not only content-oriented (or referential lexical bundles, to use Biber *et al.*'s (2004) term); disciplines may use different functional types of lexical bundles, e.g. stance bundles used as hedges are often found in social sciences, while hard sciences employ more reader-oriented bundles (Hyland, 2008). Mastering such bundles is therefore crucial to ESP or EAP learners.

In a related vein, Vašků et al. (2019) compared phraseological of-sequences in English essays by Czech novice academics, with professional academic writing. Differences in pattern use were most prominent in prepositional patterns, where novice writers overused semantically transparent patterns. Similarly, Rankin and Schiftner (2011) investigated the use of English complex prepositions by German learners. In native English, some complex prepositions have specific collocational and contextual preferences, of which the learners seemed unaware.

To conclude, corpus-informed teaching materials are potentially valuable as they contribute to learners' phraseological competence and their mastery of recurrent phraseological sequences, including discipline-specific ones. Even advanced learners tend to have a limited knowledge of phraseological sequences. Since phraseological tendencies (cf. Sinclair, 1991) pervade all levels of language, learners' insufficient phraseological competence pertains also to function word patterns such as prepositional ones. This evidence makes a case for the relevance of corpus-informed teaching materials dedicated to the phraseology of function words.

3. Material and method

The data employed in this study were drawn from corpora roughly comparable in terms of design and size: representative national corpora of English (British National Corpus, 2007) and Czech (SYN2015, Křen *et al.*, 2015, 2016), each around 100 million words. Both contain a variety of written texts; they do not entirely match as regards the time of publication. The BNC, compiled in the early 1990s, contains texts from the late 20th century (Burnard, 2009), mostly between the 1960s-1990s. The SYN2015 covers fiction and non-fiction published between 1990—2015, and journalism from 2010—2015, most texts falling under the span 2010—2014 (Cvrček and Richterová, 2020).

The composition of the English and Czech corpus roughly corresponds: the BNC represents British English and comprises 90% of written texts (fiction, journalism, academic texts, letters, essays etc.); the remaining 10% is spoken informal conversation (Burnard, 2009). By contrast, SYN2015 is written only; it contains a variety of printed and published fiction, non-fiction and journalism (Cvrček and Richterová, 2020). While aware of the two corpora not being a perfect match, their comparable size and overall nature (general representative national corpora) was the criterion for their choice.

As an initial step, a list of the ten most frequent prepositions was compiled for either language. Top ten frequent prepositions were identified manually within the frequency lists available for each corpus (Křen *et al.*, 2016 for SYN2015; and Kilgarriff, n.d. for the BNC).[2]

[2] Kilgarriff: BNC database and word frequency lists. Available from http://www.kilgarriff.co.uk/bnc-readme.html Czech National Corpus: Reference frequency lists (Srovnávací frekvenční seznamy). Available from <https://wiki.korpus.cz/doku.php/seznamy:srovnavaci_seznamy> (in Czech)

Only unambiguous prepositions were selected in Czech.[3] The choice of the English prepositions warrants a comment. Kilgarriff's wordlists were used since they are based on the entire BNC, and thus informative as to the prepositions' frequencies relative to the whole collection, showing that prepositions rank among the most frequent words in the corpus. However, the lists do not include normalised frequency information. Moreover they are based on the BNC World Edition (2001), which is no longer available, hence the frequencies differ slightly from the currently accessible XML version. On the other hand, searching for the frequencies of all prepositions in BNC XML Edition (2007), the frequency breakdown is limited to a random sample of 250,000 hits. However, the lemmatised top ten prepositions match those based on Kilgarriff, only their ranking is slightly different. Cf. Table 1.

Table 1. Top 10 English prepositions in the BNC World – as per lemmatised wordlists (Kilgarriff n.d.); compared to top 10 of a random retrievable 250,000 hit sample of preposition lemmata (BNC XML Edition, 2007).

Rank	**Prepositions in BNC World**	**Rank in whole wordlist**	**Raw freq in BNC World**	**Preposition in sample**	**Raw freq in prep sample**	**Raw freq – whole BNC XML**	**ipm – whole BNC XML**
1	of	3	3,093,444	of	59,085	3,040,670	30,928
2	in	6	1,924,315	to	50,779	2,593,740	26,382
3	to	10	1,039,323	in	36,341	1,937,966	19,712
4	for	11	887,877	for	16,925	878,741	8,938
5	on	16	680,739	with	12,748	658,584	6,698
6	with	17	675,027	on	12,524	729,558	7,420
7	at	19	534,162	at	10,092	521,697	5,306
8	by	20	517,171	by	9,893	512,215	5,210
9	from	24	434,532	from	8,393	424,972	4,322
10	as	48	201,968	as	4,304	653,610	6,648

To confirm the translation equivalence of *in* and *v*, in line with the corpus-driven (Tognini-Bonelli, 2001) approach adopted in this study, equivalents were extracted from the InterCorp v. 12 parallel corpus (Čermák and Rosen, 2012; Rosen *et al*., 2020) via the Treq application, 2.1 (Vavřín and Rosen, 2015; Škrabal and Vavřín, 2017).[4] This confirms that the prevalent English equivalent of Czech v is indeed in, see Table 2.

Table 2. English translation equivalents in InterCorp 12 as per Treq (Vavřín and Rosen, 2015).

Czech preposition	**prevalent English equivalent (*Treq*)**	**rank in SYN2015 lemmatised wordlist**	**raw freq in SYN2015**	**ipm in SYN2015**
v	in	4	2,296,562	19,075

[3] The preposition *se* (homonymous with a reflexive pronoun) was excluded. In fact, *se* ranks third in the SYN2015 wordlist (raw frequency = 3,070,434). However, a search in SYN2015 (Křen, *et al*. 2015) reveals that merely 155,508 of those instances are prepositional, the vast majority (2,306,916 hits) being the reflexive pronominal uses.

[4] The direction of translation was Czech to English, the query was lemmatised and case-insensitive. The search was performed within the entire corpus, i.e. not limited to any specific subcorpora.

As mentioned earlier, the preposition pair *in – v* was chosen due to their frequency: both *in* and *v* rank among the most frequent prepositions, as well as the most frequent words in the corpus overall (cf. Tables 1 and 2).

3.1 N-gram method – state of the art

N-gram methodology has proven a useful starting point for cross-linguistic studies working with related languages. When contrasting typologically distant language pairs such as English and Spanish, French, or Czech (Čermáková and Chlumská, 2016, 2017; Cortes, 2008; Granger, 2014; Šebestová and Malá, 2019) the methodology poses problems.

For instance, Granger (2014) compared lexical bundles in English and French across two genres (parliamentary debates and newspaper editorials), focusing on stems, i.e. combinations of subject and verb with optional pre-verbal elements (Altenberg, 1998). French was expected to employ more bundles overall. This tendency was apparent in editorials, but inconclusive in debates (ibid.: 64), indicating that phraseological tendencies may differ markedly across languages as well as registers.

Hasselgård (2017) on the other hand compared English and Norwegian 2-4-grams expressing temporal meanings. This study illustrates how n-gram methodology highlights typological differences which would be difficult to identify otherwise. The Norwegian data contained fewer recurrent n-grams overall, indicating English may have a stronger tendency towards recurrence. Yet in Norwegian, temporal n-grams formed a larger part of all the n-grams identified. Also, Norwegian n-grams corresponded to (fragments of) clauses more often (ibid.: 86). Hence, while some languages display more recurrence than others (i.e. typological properties are an important factor shaping phraseology), a language may employ phraseological means of expression to varying degrees in different semantic or functional areas, pointing towards a register-dependent distribution. Hasselgård´s study also hints towards n-gram methodology being potentially challenging even when applied to typologically related languages.

N-grams applied to the English-Czech language pair pose methodological challenges due to the typological non-correspondences. In Čermáková and Chlumská's (2016) n-gram analysis of Czech and English children's literature, English datasets yielded hundreds of n-grams, whilst in the Czech data of comparable size, only tens of n-grams were identified. This suggests that the results for each language are best examined separately as cross-linguistic comparability may be limited. In summary, previous cross-linguistic n-gram-based research indicates that typological properties and the register factor enter into a complex interplay. Further, depending on corpus design, the validity of the results is likely limited to the particular registers explored. These findings were used to inform the choice of data for the present study, namely large representative corpora, to ensure a variety of registers were represented.

In the following analysis, I use *n-gram* to refer to recurrent sequences of n words identified mechanically in corpus data, which may or may not correspond to structural units such as phrases; sometimes an n-gram comprises a complete phrase along with fragments of adjacent phrases or other structures (e.g. *of fall in love* or *fall in love with*, or *fall in love and*, where the conjunction implies a following clause; cf. Figure 1 in Section 3.2).

3.2 Engrammer software description

The data in this study was processed using the custom-made Engrammer freeware (Milička, 2019).[5] Engrammer enables searches for sequences of words of different lengths at once,

[5] *Engrammer*, available from <http://www.milicka.cz/en/engrammer/>

collapsing overlapping n-grams, e.g. *in order + in order to = in order to*. The frequencies of the individual overlapping variants can still be displayed. Figure 1 shows the Engrammer interface. The n-gram search results are in the left column. Clicking the n-gram, all variants subsumed under it are displayed in the right-hand column, together with their collocation strength and frequency. Optionally, collapsing is also available for similar n-grams ('similar' defined as differing in one position only). In Figure 1, lemmatised n-grams *fall in love with, have fall in love, to fall in love, I fall in love* etc. were collapsed. Henceforth I will be referring to the collapsed n-grams as *n-gram types* (e.g. *bear in mind, in spite of, fall in love* in Figure 1 are three different n-gram types); and individual n-gram occurrences as *n-gram tokens*.

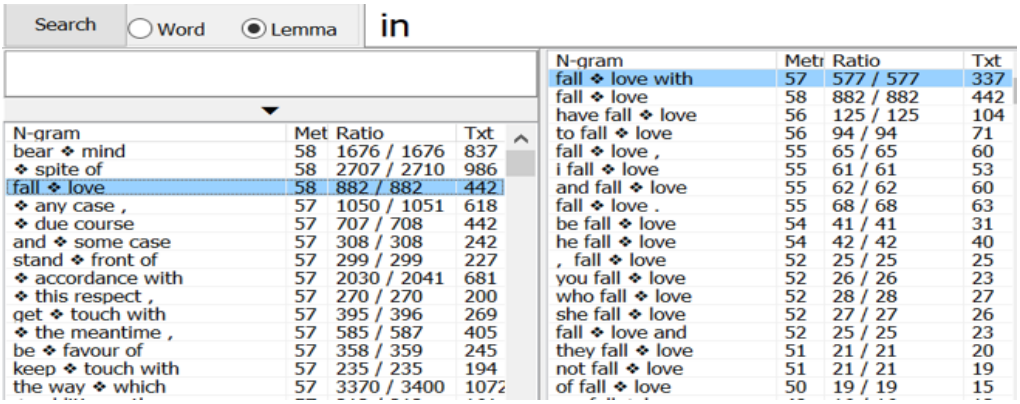

Figure 1. Engrammer interface displaying n-grams containing *in*.

3.3 N-gram search

Full text lemmatised versions of the corpora were plugged into Engrammer, one at a time. For each corpus, lemmatised 3- to 5-grams were extracted (all lengths at once), containing the preposition *in/v* in any slot (cf. Table 3). Variable word order was allowed within n-grams because Czech word order is highly flexible (cf. Čermáková and Chlumská, 2016; 2017). Given that grammatical word patterns contribute to linking, they can be expected to occur near syntactic boundaries: hence punctuation was included. The search was set so that similar n-grams (differing in one lemma only) were collapsed (cf. 3.1). The search retrieved a total of 398 n-gram types, 55,790 tokens for English; 431 n-gram types and 21,660 n-gram tokens for Czech.

Next, I analysed the collapsed n-grams manually, searching for "meaningful, linguistically structured" (Lindquist and Levin, 2008: 144) units within them, which I term *patterns*. For practical reasons, the dataset for each language was limited to the top frequent 250 (collapsed) n-gram types. Table 3 illustrates the process of identifying a pattern within lemmatised n-grams: the pattern *in front of* was abstracted from the individual n-gram types.

Table 3. Breakdown of the collapsed pattern *in front of* (span: 3-5-grams).

N-gram (lemmatised)	**N-gram token freq.**
in front of i ,	75
right in front of	89
in front of he ,	162
just in front of	70
in front of the television	69
in front of they ,	57
Total n-gram tokens	**522**
Total n-gram types	**7**

The resulting sequences were ordered by the 'risk of n-gram' rubric, using the risk ratio metric. Generally, risk ratio is based on comparing the probability of a particular item occurring in a context A as opposed to occurring in another context B (Březina, 2018: 115–16). The 'risk of n-gram' measures the strength of association between the node word (*in/v*) and each n-gram. The frequency of *in/v* in a given n-gram is compared to the frequency of *in/v* alone, and the corpus size is taken into account. E.g. *in* alone occurs 2,593,740 times in the BNC XML edition (cf. Table 1); the sequence *in front of the television* (cf. Table 3) occurs 69 times, and the corpus length is 96,986,707 tokens. This results in a risk of n-gram value of 2.1 (confidence interval = 1.8—2.2), i.e. *in front of the television* occurs at least 1.8 times more often than can be expected by chance.

While a comparable number of n-grams was extracted from both languages, English n-grams exhibited higher 'risk of n-gram' values overall than Czech (cf. Table 4), suggesting a greater degree of fixedness in English. However, this tendency may be enhanced by the analytical nature of English.

Table 4. Cross-linguistic differences in node-n-gram association strength.

English *in*		**Czech *v***	
Risk of n-gram	**No. of n-gram types**	**Risk of n-gram**	**No. of n-gram types**
57	23	52	7
56	68	51	86
55	133	50	138
54	171	49	200
Total	**395**	**Total**	**431**

3.4 Classification of *in* and *v* patterns

The prepositional patterns were sorted into functional-semantic groups in an inductive, bottom-up manner. This approach was adopted with regard to potential pedagogical applications: the most frequent patterns containing a given word can serve as the starting point for identifying the common contexts of usage of any selected word.

Where applicable, patterns were grouped based on a semantic perspective. The criterion was the meaning conveyed by lexical words in the pattern. This resulted in 6 groups of patterns, 5 of these conveying adverbial meanings. Apart from these, the *body/mind* group was singled out, since patterns referring to body parts (e.g. *go hand in hand with*) or the mind (*bear in mind*) were frequent in both corpora.

Since not all patterns lend themselves to semantic classification, the semantic perspective was complemented with a formal-structural one wherever no overarching semantic feature was identified, but multiple patterns shared a grammatical structure or part of speech: e.g. complex preposition patterns (*in front of, v rámci* 'in the framework of'[6]), or patterns comprising a 'copula + complement' (*be in charge, být v pořádku* 'be in order').

Finally, two groups of patterns stood out: patterns conveying emphasis (*in the first place, v prvé řadě* 'in the first place') and hedging patterns (*in a sense, v jistém smyslu* 'in a sense'). Both were subsumed under the broadly conceived 'pragmatic' patterns, defined by fulfilling a discourse function, rather than by semantics or formal characteristics.

Some patterns could be classified by more than one of the three types of criteria (semantic, formal-structural, pragmatic): e.g. *v žádném případě* 'in no case/by no means' or *v mnoha ohledech* 'in many respects' could be classified semantically as adverbial patterns of regard, or pragmatically as emphasizers. The semantic criterion was prioritised and the patterns were classified as adverbial, since the adverbial group was considered broader and able to

[6] Henceforth, all verbatim translations from Czech into English, given in single quotation marks, are mine.

encompass the pragmatically specialized usages. Similarly, wherever a pattern conveyed adverbial meaning but also contained a distinctive structural element, e.g. a complex preposition or phrasal verb (e.g. *ve srovnání s rokem X* 'in comparison with the year X'), it was classified under the corresponding structural pattern type rather than the semantic adverbial type, in line with the focus on phraseological patterning centred around function words.

3.5 Idiomaticity as an additional criterion

Independently of the classification based on semantic/formal/functional criteria, I annotated the patterns for idiomaticity, defined broadly as being lexically (at least partly) fixed: either a given word cannot be replaced with its (near) synonym: e.g. *be in short supply* not **be in brief/abbreviated supply*; or the choice of acceptable synonyms is limited: *be not in a position;* possibly also *be not in a place*[7], but not **be not in a location*.[8]

Idiomatic patterns occurred across the pattern groups and will be discussed in Section 4.6. The decision to add this perspective was prompted by the occurrence of potentially metaphorical patterns among the *body/mind* pattern group, e.g. *hand in hand* (cf. 6.2). Next, idiomatic patterns were assessed in terms of semantic transparency/opacity. Patterns were considered opaque if the whole pattern conveyed a meaning which was not a sum of the meanings of its parts (e.g. *in the light of these*), their meaning was perceived as figurative rather than literal (*keep in touch with*), or they contained a limited-collocability item (*in the nick of*). As shown by Table 5, the proportion of transparent and opaque patterns is even in both languages; idiomatic patterns were more frequent in English overall.[9]

Table 5. Idiomatic patterns in English and Czech.

Fixed patterns	**English**	**Czech**
Opaque	26	16
Transparent	22	15
Total – idiomatic patterns	**48**	**31**
Total – all patterns	**250**	**250**

A variety of meanings and functions is conveyed by *in* and *v* patterns. Table 5 outlines the pattern groups identified, ordered by frequency for each language corpus, listing pattern type frequencies for each group.[10] Most pattern groups were identified in both English and Czech. Pattern groups identified in one language only are addressed in Section 5.

Table 6 lists the pattern groups according to their respective defining criteria: structural, semantic or pragmatic. Section 4 goes on to discuss the attested pattern groups.

[7] One example was attested in the BNC: *Thee ain't in no place to talk about prying*; possibly informed by analogy with *it is not my place to*.

[8] This can be viewed as a manifestation of Sinclair's (1991:110) principle of idiom, i.e. "a large number of semi-preconstructed phrases that constitute single choices", or as Altenberg (1998: 115) puts it "more or less prefabricated or routinized building blocks".

[9] Admittedly it proved difficult to establish robust criteria for determining semantic opacity. A possible solution would be having the patterns evaluated by native speakers, followed by an inter-rater agreement analysis.

[10] E.g. complex preposition patterns comprised 44 pattern types, one of them being *in front of* (described in Table 3). Higher pattern type frequency indicates a greater formal variety within the particular pattern group. Contrarily, a low pattern type frequency points towards a greater degree of formal repetitiveness within that group.

Table 6. Pattern groups in English and Czech.

Pattern group	English		Czech	
	Example of pattern type	**Pattern type freq**	**Example of pattern type**	**Pattern type freq**
Structural				
Complex prep.	in front of	44	v rámci NP 'within the framework of NP'	28
Complex conj.	in order to	21	N/A	0
Copular/phasal verb	be in charge	20	být v pořádku 'be in order / all right' pokračovat v chůzi 'continue walking'	20
Phrasal/prep. verb	come in handy	10	spočívat v tom, že 'lie/consist in the fact that'	28
Valency	interested in	7	N/A	0
Semantic				
ADV place	in chapter…, in appendix…	33	pobyt v nemocnici 'a stay in hospital'	46
ADV regard	and in some case	23	v tomto ohledu 'in this respect'	19
ADV manner	way in which	22	ve zkratce 'in short'	1
ADV time	in the morning	18	aktivní v noci 'active at night	44
ADV circumstances/ state	in silence, in doubt	9	přednost v jízdě 'right of way' být v klidu 'be calm'	23
Body/mind	in a ADJ voice	4	sucho v ústech 'dryness in the mouth'	19
Pragmatic				
Emphasis	in the first place, in any case	35	v prvé/neposlední řadě 'in the first place'/'last but not least'	17
Hedge / approximation	in a sense	4	v jistém smyslu být 'in a sense be'	4
Other	N/A	0	minulý měsíc ubývat v 'last month decrease in'	1
TOTAL		**250**		**250**

4. Discussion of pattern uses

4.1 Semantically defined patterns: Adverbial patterns

This group includes patterns expressing adverbial meanings, as illustrated by examples (1) through (5).

(1) Place: *in* court / sedět *v* kuchyni 'be sitting in the kitchen'

(2) Time: *in* the morning / aktivní *v* noci ‘active at night’

(3) Manner: *in* short / *ve* zkratce ‘in short’

(4) Regard: *in* this respect / *v* tomto ohledu ‘in this respect’

(5) State: if *in* doubt / být *v* klidu ‘be calm’

Some state adverbial patterns could form part of copular constructions; yet the n-grams retrieved did not contain the copula, e.g. (*být*) *jako v bavlnce* ‘(to be) comfortable’.

4.2 Semantically defined patterns: Body/mind patterns

Patterns containing a noun referring to body parts or the mind, see example (6), were singled out; idiomaticity was taken into account as a result, since these expressions are frequent source domains for metaphors (Lindquist and Levin, 2008).

(6) hand in hand / říci si v duchu ‘say to oneself’

4.3 Structurally defined patterns: Verbal patterns

In verbal patterns, copular (example 7), phrasal and prepositional (8) verbs occurred. This was not surprising since all these verbs form part of phraseological sequences: copular verbs require complementation, while phrasal/prepositional verbs constitute multi-word units by definition.

(7) be in charge / být v pořádku ‘be in order/all right’

(8) come in handy/ spočívat v tom, že ‘consist in the fact that’

One verbal pattern group was limited to English: verbs with a valency complement, e.g. *interested in*. These are discussed in Section 5.

4.4 Structural: Complex prepositions and conjunctions

Another group of patterns was formed by complex prepositions (9) and conjunctions (10), the latter only attested in English.

(9) in front of /v rámci NP ‘within NP’

(10) in order to

4.5 Pragmatic patterns

Lastly, patterns with pragmatic functions were identified: emphasis (example 11) and hedging/approximation (12).

(11) Emphasis: in the first place / v neposlední řadě

(12) Hedge: in a sense / v jistém smyslu

While some functional-semantic pattern groups comprise a diverse set of expressions (e.g. place adverbials), the pragmatic group seemed limited to few recurrent patterns. This suggests that the pragmatic functions may favour more conventionalised forms of realization.

4.6 Idiomatic patterns

Examples of idiomatic patterns were found across a range of semantically/structurally/pragmatically defined pattern groups, as shown in Tables 7 and 8 in decreasing order of frequency for each language.

Table 7. English idiomatic patterns: distribution across pattern groups.

Group	Opaque	Transparent	Total types
Copular/phrasal	12	3	15
Emphasis	3	5	8
Complex prep.	4	3	7
Phrasal/prep verb	1	4	5
Time	3	2	5
Body/mind	2	2	4
Circumstances/state	0	3	3
Valency	1	0	1
Total types	**26**	**22**	**48**

In English, most idiomatic patterns occurred in the verbal type comprising a copular or phrasal verb (13), followed by patterns, serving to emphasize, structure and punctuate discourse (14); and complex prepositions, likewise means of text structuring (15).

(13) fall in love; get in touch with

(14) in any case; in the first place

(15) in spite of; in the wake of the

Table 8. Czech idiomatic patterns: distribution across pattern groups.

Group	Opaque	Transparent	Total types
Body/mind	3	8	11
Copular/phrasal	4	4	8
Circumstances/state	4	1	5
Place	3	0	3
Phrasal/prep. verb	2	1	3
Manner	0	1	1
Total types	**16**	**15**	**31**

Among Czech idiomatic patterns, especially those referring to body and mind were prominent (16), followed by copular verbal patterns (17).

(16) jít ruku v ruce ‘go hand in hand’

(17) být v sedmém nebi ‘be in seventh heaven’

Notably, some adverbial circumstances/state patterns potentially overlap with verbal ones: the pattern in ex. (18) would often occur with the copula *být* (‘be’). However, the copula was not included in the recurrent pattern since it can alternate with other verbs. Ex. (19) could be alternatively classified under *phrasal/prepositional verb* (cf. 21 below).

(18) (být) jako v bavlnce ‘be very comfortable’

(19) nechat ve štychu ‘leave in the lurch’

5. Cross-linguistic differences

The first major cross-linguistic difference lies in the distribution of pattern groups. Essentially, *in* and *v* patterns convey the same functions in both languages. However, pattern types are distributed differently: ranked by raw frequency, corresponding pattern groups differ in their position within the top-frequent ranking (see Table 9). In other words, Czech does not employ individual pattern types with the same frequency as English. To illustrate this, Table 9 lists the top frequent five pattern groups in both languages. The frequency was assessed by the n-gram token counts within each pattern group, i.e. by the number of all n-grams conveying this function. The patterns in *italics* (place adverbials, complex prepositions) rank among the top five in both languages.

Table 9. Top five functions for each language, ordered by n-gram token frequency.

English	**Token freq.**	**Czech**	**Token freq.**
Complex prep.	11,430	Time	8,373
Emphasis	7,183	Phrasal / prep. verb	5,722
Manner	4,375	*Complex prep.*	5,420
Place	4,239	*Place*	5,376
Copular	3,853	Regard	4,172

Secondly, two pattern groups were identified in English only, highlighting its analytic features: complex conjunctions, and verbs with valency complements. Below I discuss the language-specific features revealed by the pattern analysis for each language.

5.1 English *in* patterns

There were two extra pattern groups attested in English: complex conjunctions and prepositional verbs with valency complements. As regards complex conjunctions, the majority of this group was represented by *in order to* or variations thereof (18 out of 21 n-gram types). Either there is an adjectival head postmodified by an infinitival clause introduced by *in order to* (example 20); or the pattern captures the following verb (example 21). The other 2 conjunction pattern types were *in such a way as/that* and *except in so far*.

(20) necessary in order to

(21) in order to achieve/gain/understand/avoid

Since the English and Czech corpora did not entirely match in terms of the text types represented (cf. Section 2), the question arises whether complex conjunctions may be limited to English due to their distribution in specific text types, perhaps less represented in the Czech corpus. This was checked for the most frequent conjunction *in order to*. As apparent from Table 10, *in order to* is predominantly found in books; a closer inquiry into its distribution across text domains shows that it occurs predominantly in social sciences, followed by world affairs (i.e. newspapers). Interestingly, *in order to* is widely used in social sciences (215 ipm) while much less common in natural sciences (143 ipm). This evokes Hyland's (2008) findings about specialized discourses being marked by the usage of text-structuring patterns.

Table 10. *in order to* - distribution across text types in BNC.

Text type	No. of words	Freq .raw	Freq. ipm
Written books and periodicals	79,187,792	10,243	129.35
Written miscellaneous	7,437,161	1,292	173.72
Context-governed	6,175,896	485	78.53
Demographically sampled	4,233,962	16	3.78
Written-to-be-spoken	1,278,618	14	1.95
Total	**98,313,429**	**12,050**	**122.57**

Furthermore, complex prepositions are more frequent in English (44 n-gram types, 11,430 n-gram tokens) than in Czech (28 n-gram types, 5,420 n-gram tokens). Not only are complex preposition patterns almost twice as frequent in English overall (cf. n-gram-token counts), they are also formally more varied (= more n-gram types). In sum, the findings about complex conjunctions and prepositions indicate there may be more complex function patterns in English overall, in line with English being an analytic language with a rich and recurrent repertoire of function words.

The second type of pattern exclusive to English was a verb followed by its valency complement; a prepositional object (22) or adverbial prepositional phrase (23).

(22) interested in NP

(23) an increase in NP

Although this group was not attested in Czech, there were similar Czech patterns, namely a verb followed by a prepositional phrase, as in (24).

(24) vzít v potaz/úvahu ‘take into account/consideration’

However, in Czech, the noun phrase complementing the preposition is lexically fixed. Czech patterns such as *vzít v* + NP are idiomatic, hence the choice of the noun *potaz/úvahu* (‘take into account’)*;* while in English patterns as in *interested in*, the following slot is open and may contain any one of a range of nominal complements. Due to this collocational fixedness, Czech patterns of the type *vzít v úvahu* were labelled as phrasal/prepositional verb. (Admittedly such Czech constructions are not formally analogous to English phrasal verbs; yet they are characterised by lexical fixedness).

A glimpse at the n-gram type-token ratios of the attested patterns reveals that some pattern groups are formally repetitive; a qualitative look at the data confirms this. English adverbial patterns of manner consist almost exclusively of *the/a way in which* (18 of total 22 n-gram types). A similar tendency was observed in Czech adverbial patterns of regard (variations on *v tomto případě* ‘in this case’, *v tomto ohledu* ‘in this respect’). Complex prepositions are likewise repetitive in both languages, which can be expected given their formal fixedness.

Similarly, body/mind patterns comprise a mere four types – cf. Table 11 (or rather three, given the overlap *go hand in hand with*). Despite its repetitiveness, the body/mind pattern group is very frequent: it would warrant closer investigation to find out more about its common contexts of use.

Table 11. Body/mind pattern group.

N-gram	Collocation strength (risk of n-gram)	Freq.
bear in mind	57.012	1676
go hand in	56.7823	172
hand in hand with	56.3167	114
say in a low voice	55.1806	62

5.2 Czech *v* patterns

Similarly to English, an examination of type-token ratios provides some insights into Czech prepositional patterns. Place adverbial patterns are a diverse group comprising a number of given names, whose referents range from TV series (25) to institutions (26) or even topical events (27).

(25) Ordinace v růžové zahradě 'Surgery in the Rose Garden' [11]
Sex ve městě 'Sex and the City'

(26) fakulta UK v Praze 'faculty of Charles University in Prague'
krajský soud v Brně 'regional court in Brno'

(27) olympiáda v Soči 'Olympics in Sochi'

Other adverbial place patterns are register-specific, as in (28), typical of the language of advertising.

(28) info o ceně v obchodě – 'price information available in the shop'

Further, place patterns refer to a variety of locations (29). Indeed, *v* is one of the most common prepositions to combine with the locative (Cvrček *et al.*, 2015: 172).[12]

(29) v nemocnici / v kuchyni / ve vězení – 'in hospital/the kitchen/prison'

Lastly, idiomatic place patterns were found (30).

(30) viset ve vzduchu 'hang in the air'; praskat ve švech 'burst at the seams'

On the other end of the diversity cline are pragmatic patterns expressing emphasis, mostly variations on *v žádném/každém případě* 'by no/all means'. This may reflect the tendency of pragmaticalised patterns to become fixed with repeated usage. By contrast, adverbial place patterns may refer to a host of referents, reflecting speakers´diverse communicative needs.

Finally, more idiomatic patterns were attested in English than in Czech overall. A qualitative assessment of the idiomatic patterns seems to suggest that there is in fact a cline of semantic opacity, as illustrated by (31–33) (note: 31 and 32 are equivalents which occured in both languages).

(31) fully opaque, non-compositional: be in full swing - být v plném proudu

(32) abstract uses (e.g. personifications): go hand in hand – jít ruku v ruce s

(33) fully transparent: put in an appearance – být jako v transu 'be as if in a trance'

[11] A popular Czech soap opera.
[12] My thanks go to the anonymous reviewer for pointing this out.

6. Conclusion

This pilot study has examined prepositional patterns in English and Czech, classifying them into inductively defined groups based on semantic, structural or pragmatic criteria. Major pattern types represented in both languages included adverbial patterns, verbal patterns, complex prepositions, and conjunctions. Pragmatic patterns served as a means of emphasis or hedging. While the pattern groups generally corresponded between the two languages, they are distributed differently: e.g. complex prepositions occurred nearly twice as often in English than in Czech. The distribution may be influenced by text type or register – more research into this is needed. A potential application of this finding would present itself in the use of custom-made corpora of specialized texts in the classroom, enabling students to identify patterns and compare their uses in their L1 and L2, or to observe whether translation equivalent patterns are used in similar contexts or registers.

To some extent, patterns reflected the typological properties of the languages. Analytical English employs more complex prepositions and conjunctions, both in terms of n-gram type and token counts. As earlier reseach has indicated that even advanced EFL learners may tend to use fewer patterns and prefer less lexically sophisticated ones (Vašků *et al*., 2019), this is further evidence that the use of such complex text-structuring patterns deserves attention in class.

Finally, the pattern types display a varying degree of repetitiveness. This may be caused by some meanings being more closely associated with particular expressions (*v žádném případě* – 'under no circumstances'). Alternatively, it may simply reflect the high frequency of some patterns in the corpus (*in order to*). These hypotheses prompted by the pilot study findings provide an interesting impetus for further research; the reasons for the differences in individual patterns' frequencies could be investigated through a qualitative analysis of a larger dataset. At any rate, the observations regarding pattern idiomaticity suggest that this parameter warrants special attention in language instruction. Under an inductive teaching approach, similar observations about specific patterns and their usage can be made efficiently by students exploring corpus data.

To complement this study, patterns around other frequent prepositions should be compared to *in* and *v*. Lastly, bearing in mind that phraseological patterns can identify register-specific features (Biber *et al*., 2004), another follow-up possibility is a comparison of the prepositional patterns identified in large representative general corpora such as the BNC and SYN2015, to patterns found in specialized corpora of particular registers – building on research on register variation in Czech (Cvrček *et al*., 2020).

The results of the study have illustrated the potential value of viewing phraseological sequences through a cross-linguistic lens: contrasting prepositional patterns in two corpora of different languages reveals similarities in the pattern types employed by the languages, while also highlighting differences in the distribution, overall frequency, functional load and diversity of pattern types. Given the importance of phraseological competence for L2 proficiency (Paquot, 2018), contrastive phraseological analyses can provide advanced learners with valuable insight into their target language phraseologies. Further, patterns may illustrate the typological features of languages, as reflected in the greater frequency of complex prepositions and conjunctions in analytical English – such observations may help L2 learners better grasp the theoretical notion of language typology as well as to notice structural differences between languages.

Acknowledgements

My sincere thanks go to Jiří Milička for methodological assistance, as well as to the reviewers for their thorough reading and helpful critical comments.

References

Altenberg, B. 1998. On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In *Phraseology: Theory, Analysis and Applications*, A.P. Cowie (ed.), 101–22. Oxford: OUP.

Biber, D, Conrad, S. and Cortes, V. 2004. 'If You Look at…': Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25(3):371–405. doi: 10.1093/applin/25.3.371.

Březina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: CUP.

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Burnard, Lou. 2009. What Is the BNC? Oxford Text Archive, IT Services, University of Oxford. Available at http://www.natcorp.ox.ac.uk/corpus/index.xml [Last accessed 21 April 2021].

Čermák, F. and Rosen, A. 2012. The Case of InterCorp, a Multilingual Parallel Corpus. *International Journal of Corpus Linguistics* 13(3):411–27.

Čermáková, A. and Chlumská, L. 2016. Jazyk dětské literatury: kontrastivní srovnání angličtiny a češtiny. In *Jazykové paralely*, A. Čermáková, L. Chlumská and M. Malá (eds), 162–187. Prague: NLN.

Čermáková, A. and Chlumská, L. 2017. Expressing Place in Children's Literature. Testing the Limits of the N-Gram Method in Contrastive Linguistics. In *Cross-linguistic Correspondences: From Lexis to Genre*, *Studies in Language Companion Series*, T. Egan and H. Dirdal (eds), 75–95. Amsterdam: John Benjamins.

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

Cvrček, V. *et al*. 2015. *Mluvnice současné češtiny 1: Jak se píše a jak se mluví*. Charles University, Karolinum.

Cvrček, V., Laubeová, Z., Lukeš, D., Poukarová, P., Řehořková, A. and Zasina, A.J. 2020. *Registry v češtině*. Praha: NLN.

Cvrček, V. and Richterová, O. (eds) 2020. *En:Cnk:Syn2015*. Available at http://wiki.korpus.cz/doku.php?id=en:cnk:syn2015&rev=1598975168 [Last accessed 15 August 2021].

Czech National Corpus: Reference frequency lists (Srovnávací frekvenční seznamy). 2016. Institute of the Czech National Corpus. Available from: http://www.korpus.cz

Ebeling, S.O. and Hasselgård, H. 2015. Learner Corpora and Phraseology. In *The Cambridge Handbook of Learner Corpus Research*, S. Granger, G. Gilquin and F. Meunier (eds), 207–230. Cambridge: CUP.

Erman, B. and Warren, B. 2000. The Idiom Principle and the Open Choice Principle. *Text* 20(1):29–62.

Granger, S. 2014. A Lexical Bundle Approach to Comparing Languages: Stems in English and French. Special issue of *Languages in Contrast* 14(1), M.-A. Lefer and S. Vogeleer (eds), 58–72. doi: 10.1075/lic.14.1.04gra.

Granger, S. 2017. Academic Phraseology. A Key Ingredient in Successful L2 Academic Literacy. *Oslo Studies in Language* 9(3), *Academic Language in a Nordic Setting – Linguistic and Educational Perspectives*, R.V. Fjeld, K. Hagen, B. Henriksen, S. Johansson, S. Olsen and J. Prentice (eds), 9–27.

Granger, S. 2018. Tracking the Third Code. In *The Corpus Linguistic Discourse: In Honour of Wolfgang Teubert*, *Studies in Corpus Linguistics*, A. Čermáková and M. Mahlberg (eds), 185–204. Amsterdam: Benjamins.

Granger, S. and Bestgen, Y. 2014. The Use of Collocations by Intermediate vs. Advanced Nonnative Writers: A Bigram-Based Study. *International Review of Applied Linguistics in Language Teaching (IRAL)* 52(3):229–52.

Granger, S. and Meunier, F. (eds). 2008. *Phraseology. An Interdisciplinary Perspective*. Vol. 139. Amsterdam: Benjamins.

Granger, S. and Paquot, M. 2008. Disentangling the Phraseological Web. In *Phraseology:An Interdisciplinary Perspective*, S. Granger and F. Meunier (eds), 27–49. Amsterdam; Philadelphia: Benjamins.

Gray, B. and Biber, D. 2015. Phraseology. In *The Cambridge Handbook of English Corpus Linguistics*, D. Biber and R. Reppen (eds), 125–145. Cambridge: CUP.

Groom, N. 2010. Closed-Class Keywords and Corpus-Driven Discourse Analysis. In *Keyness in Texts*, M. Bondi and M. Scott (eds), 59–78. Amsterdam: Benjamins.

Hasselgård, H. 2017. Temporal Expressions in English and Norwegian. In *Contrasting English and other Languages through Corpora*, M. Janebová, E. Lapshinova-Koltunski and M. Martinková (eds), 75–101. Newcastle: Cambridge Scholars Publishing.

Hasselgård, H. 2019. Phraseological Teddy Bears: Frequent Lexical Bundles in Academic Writing by Norwegian Learners and Native Speakers of English. In *Corpus Linguistics, Context and Culture*, V. Wiegand and M. Mahlberg (eds), 339–362. Berlin, Boston: De Gruyter.

Hasselgren, A. 1994. Lexical Teddy Bears and Advanced Learners: A Study into the Ways Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics* 4(2):237–59. doi: https://doi.org/10.1111/j.1473-4192.1994.tb00065.x.

Howarth, P. 1998. Phraseology and Second Language Proficiency. *Applied Linguistics* 19(1):24–44. doi: https://doi.org/10.1093/applin/19.1.24.

Huang, L.-F. 2011. *Discourse Markers in Spoken English: A Corpus Study of Native Speakers and Chinese Non-Native Speakers*. Doctoral thesis, University of Birmingham, Birmingham, UK.

Hunston, S. 2008. Starting with the Small Words. Special issue of *International Journal of Corpus Linguistics* 13(3): *Patterns, Meaningful Units and Specialized Discourses*, U. Römer and R. Schulze, 71–95.

Hyland, K. 2008. 'As Can Be Seen': Lexical Bundles and Disciplinary Variation'. *English for Specific Purposes* 27:4–21.

Kilgarriff, A. n.d. BNC Database and Word Frequency Lists. Available at https://www.kilgarriff.co.uk/bnc-readme.html [Last accessed 25 June 2021].

Klégr, A. and Malá, M. 2009. English Equivalents of the Most Frequent Czech Prepositions: A Contrastive Corpus-Based Study. In *Proceedings of the Corpus Linguistics Conference, CL 2009, Conference in Liverpool, 20-23 July 2009*, M. Mahlberg, V. González-Díaz and C. Smith (eds). Available at http://ucrel.lancs.ac.uk/publications/cl2009/ [Last accessed 26 June 2021].

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. and Zasina, A. 2015. *SYN2015: reprezentativní korpus psané češtiny*.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. and Zasina, A. 2016. SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2522–2528. Portorož: ELRA.

Lindquist, H, and Levin, M. 2008. Foot and Mouth: The Phrasal Patterns of Two Frequent Nouns. In *Phraseology. An Interdisciplinary Perspective*, S. Granger and F. Meunier (eds), 143–158. Amsterdam: Benjamins.

Milička, J. 2019. *Engrammer*. Praha: Institute of the Czech National Corpus, Faculty of Arts, Charles University.

Paquot, M. 2018. Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly* 15(1):29–43. doi: https://doi.org/10.1080/15434303.2017.1405421.

Paquot, M. and Granger, S. 2012. Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics* 32:130–49. doi: 10.1017/S0267190512000098.

Peřestá, G. 2017. *Akviziční interference angličtiny a češtiny v prepozicionálních konstrukcích / Acquisitional Interference of English and Czech in Prepositional Constructions*. Unpublished BA thesis, Charles University, Faculty of Arts, Prague.

Rankin, T. and Schiftner, B. 2011. Marginal Prepositions in Learner English: Applying Local Corpus Data. Special issue of *International Journal of Corpus Linguistics* 16(3), *Applying Corpus Linguistics*, F. Farr and A. O'Keeffe (eds), 412–434.

Reppen, R. 2011. Using Corpora in the Language Classroom. In *Materials Development in Language Teaching*, B. Tomlinson (ed.), 35–50. Cambridge: CUP.

Rosen, A., Vavřín, M. and A.J. Zasina. 2020. *The InterCorp Corpus*, Version 13 of 1 November 2020.

Šebestová, D., and Malá, M. 2019. Expressing Time in English and Czech Childrens Literature: A Contrastive N-Gram-Based Study of Typologically Distant Languages' In *Language Use and Linguistic Structure: Proceedings of the Olomouc Linguistics Colloquium 2018*, 469–483 Olomouc: Palacky University.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Škrabal, M. and Vavřín, M. 2017. Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii* 99(2):245–60.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: Benjamins.

Vašků, K., Brůhová, G. and Šebestová, D. 2019. Phraseological Sequences Ending in of in L2 Novice Academic Writing. In *Computational and Corpus-Based Phraseology. EUROPHRAS 2019, Lecture Notes in Computer Science*, G. Corpas Pastor and R. Mitkov (eds), 431–443. Cham: Springer.

Vavřín, M. and Rosen, A.. 2015. Treq (v. 2.1). *Treq*. https://treq.korpus.cz/ [Last accessed 28 April 2021].

Author's address

Denisa Šebestová
Faculty of Arts, Department of English Language and ELT Methodology
Charles University
nám. Jana Palacha 1/2
CZ-Prague 110 00
Czech Republic
denisa.sebestova@ff.cuni.cz

TELLING in English, Norwegian and French: A three-way contrast

Thomas Egan

Inland Norway University of Applied Sciences (Norway)

This paper presents the results of a study of double object constructions containing the cognate verbs English *tell* and Norwegian *fortelle*, based on data from the English–Norwegian Parallel Corpus. The results show that there is a certain degree of correspondence between the two verbs in constructions with nominal direct objects, with less mutual correspondence in constructions with finite clausal objects, very little correspondence in constructions with objects in the form of direct speech, and none whatsoever in the case of non-finite clausal objects, which only occur with *tell*. The paper then expands the topic to include TELL predications in French. The data were retrieved from the Oslo Multilingual Corpus. It transpires that the form of French translations of Norwegian expressions are more similar, at least for some constructions, to the Norwegian originals than are their English counterparts.

Keywords: ditransitives, cognates, TELL verbs, SAY verbs, direct speech, English/French/Norwegian

1. Introduction

This paper is divided into two parts. The first part presents the results of a study of double object constructions containing the cognate verbs English *tell* and Norwegian *fortelle*, based on data from the English–Norwegian Parallel Corpus (ENPC: see Johansson, 2007: 10). It is part of a larger study of a handful of cognate verbs, coding actions of GIVING, SENDING, BRINGING, LENDING and SELLING, as well as TELLING. The English verb *tell* is the second most frequent ditransitive verb in English, after *give* (Mukherjee, 2005: 119), as well as the second most frequent communication verb, after *say* (Biber *et al*., 1999: 368; Viberg, 1996: 156). It differs from most other English communication verbs in occurring in the ditransitive construction (Huddleston and Pullum, 2002: 310). The second part of the paper examines translations into both English and French of Norwegian TELL predications in the Oslo Multilingual Corpus (OMC: see Johansson, 2007: 18). The reason for including French in the study is that it resembles Norwegian, but not English, in containing a SAY verb (*dire*) that partakes of the dative alternation.

An analysis of GIVE constructions in English and Norwegian shows that these are remarkably similar, both in their semantics and their distribution (Egan, forthcoming). The distribution of the ditransitive and prepositional dative constructions in the two languages in the ENPC is actually more similar than it is between the different varieties of spoken English

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.v11i1.3438.

analysed by Szmrecsanyi *et al.* (2017). Moreover, the semantic network of the English and Norwegian verbs is almost identical, in terms of both central and peripheral senses. The only difference of note is a greater tendency for *give* to occur in light verb constructions (*give a kiss/glance/push* etc.).

The TELL verbs in the two languages differ from the GIVE verbs in at least one important respect. As pointed out by Mukherjee (2005: 127), the direct object is much more likely to take the form of a clause, as in (1) and (2).

(1) He did once tell me *that he hated shaking hands*. (RDA1)[1]
Han fortalte meg en gang *at han hatet å håndhilse*. (RDA1T)

(2) One time he told me *what the name of the town meant*. (NG1)
En gang fortalte han meg *hva navnet på byen betydde*. (NG1T)

The constructions in the translations of the *that*-clause in (1) and the *wh*-clause in (2) mirror those of the originals.[2] In this paper the following three research questions are addressed.

1. How similar to/different from one another are the distributions of double object constructions containing the verbs *tell* and *fortelle* in the original texts in English and Norwegian?

2. Are there some kinds of tokens that are either usually or never translated by congruent constructions? If never, what characterises the divergent translations?

3. What are the French translation correspondences of the English and Norwegian constructions?

The first of these research questions is answered by comparing the source texts in English and Norwegian, the second by comparing the target texts in Norwegian and English with their sources, and the third by comparing French and English translations to one another and to their Norwegian sources. As for the structure of this paper, section 2 presents the corpus data and the methods employed to analyse them. Section 3 compares English and Norwegian with respect to constructions containing various types of direct objects. Section 4 expands the topic to include TELL predications in French, and finally, section 5 contains a summary and conclusion.

2. Theory, corpus and method

The reason for studying cognate verbs that occur in identical syntactic constructions is grounded in the assumption that translators, in addition to attempting to render the semantic and pragmatic import of their source texts, will tend to employ congruent constructions where these are available in the target language (see Ebeling, 1998: 169). Moreover, cognates tend to trigger cognates in the mental lexicon of bilingual speakers (Paradis, 2004: 218, Vandevoorde, 2020: 205–209). When a cognate lexeme can be used in an equivalent grammatical construction in a target language, one might expect translators to choose to employ them both.

[1] The first part of the code 'RDA1' refers to the text in the English–Norwegian Parallel Corpus from which the example has been taken, with 'RDA' being the initials of the Egan. 'RDA1T' stands for the translation of the same text. The full titles of the original works and the translations are listed in Johansson (2007: 329-338).
[2] The labels '*that*-clause' and '*wh*-clause' will be used throughout for subordinate declarative and interrogative clauses, respectively, in all three languages.

Numerous papers have been published on double object constructions in English. Mukherjee (2005: 3−63) contains a comprehensive overview of ditransitive constructions, and recent years have seen the publication of multifactorial studies of the English dative alternation by Bresnan and Hay (2008), Bresnan and Ford (2010), Szmrecsanyi *et al*. (2017), Röthlisberger *et al.* (2017), among others. Much less has been written about these constructions in Norwegian: among those who have addressed them are Åfarli (1992), Brøseth (1998), Tungseth (2008) and Lohndal (2011). Andersen *et al*. (2012: 24) state that "the DA [dative alternation] in Norwegian is very similar to that in English, at least in the most straightforward cases". As mentioned above, Egan (forthcoming) shows considerable similarity in the case of the prototypical GIVE verbs in the two languages.

Since one of the aims of the present study is to compare the two types of double object constructions, the ditransitive and the prepositional dative, the data investigated are limited to active voice examples with an explicitly coded TELLER (except in the case of imperatives) and explicitly coded THEMEs, encoded syntactically as direct objects, and RECIPIENTs, encoded syntactically as indirect or prepositional objects. Thus, examples of monotransitive TELL constructions in the source texts are excluded from consideration. Also excluded are constructions of the type labelled 'indirect object + prepositional object' by Quirk *et al*. (1985: 1208), as in 'tell x about y', 'tell about' being considered a prepositional verb, as in Mukherjee (2005: 126). Examples with passive verbs, in which either the THEME or the RECIPIENT is encoded as a syntactic subject rather than an object, are also not included, since these normally contain just two participants.

Two corpora are used in the present study. The initial comparison of English and Norwegian is based on data from the ENPC, which contains extracts from 50 English texts, both fictional and non-fictional, aligned with their translations into Norwegian, and extracts from 50 texts in Norwegian with their English translations. These extracts are between 10,000 and 15,000 words in length, yielding a total of about 650,000 words of both original text in, and translations into, each language. Although the corpus is rather small, the facts that it is bidirectional and that the two lexemes in the study are both relatively common renders it suitable for the present study. All tokens containing forms of the lemmas *tell* and *fortelle* in the original texts were extracted from the corpus. For English, the forms are *tell*, *tells*, *telling* and *told*. For Norwegian, the forms are *fortell* (imperative), *fortelle* (infinitive), *forteller* (present), *fortalte* (preterite) and *fortalt* (past participle). The tokens retrieved were sorted manually to only include all instances with an explicit subject (except in the case of imperatives) and two explicit objects. The direct object in English may be a finite clause, as in (1) and (2), a non-finite clause, as in (3), or an NP, as in (4). It may also consist of direct speech, which in both written English and Norwegian is enclosed in quotation marks, as in (5)−(7). When the direct object takes the form of a clause or direct speech, the RECIPIENT is always encoded by an indirect object, never a prepositional one (Levin, 1993: 203).

(3) Then she told Mum *to leave*. (BO1)[3]
Så *ba* hun mamma gå. (BO1T)
Then asked she Mum (to) leave.[4]

[3] Note that Huddleston and Pullum (2002: 1207) would analyse the non-finite clause in (3) as a catenative complement rather than a direct object. There is also a construction in which a *to*-infinitive clause is preceded by a *wh*-word, as in 'x told y wh. *to*-infinitive'. Huddleston and Pullum (2002: 1264) point out that this sort of clause resembles a finite *wh*-clause in its distribution. I have classified examples of this construction as *wh*-clauses.

[4] An English gloss is provided in italics for the relevant part of the predication in Norwegian whenever this is not faithfully rendered by the English translation in the corpus, or when it is not a faithful rendition of an English original.

(4) I'll tell you *a story*. (OS1)
Jeg skal fortelle deg *en historie*. (OS1T)

(5) He told her, deadpan, *"Love can happen to the elderly, too."* (AH1)
Han sa gravalvorlig: *"Kjærligheten kan komme til gamlinger også."* (AH1T)
He said, serious as the grave, "Love can come to oldies too".

(6) *"The King of England uses only five inches of bath water,"* Aunt would tell them. (AB1)
"Kongen av England bruker bare fem tommer med badevann," fortalte tante dem. (AB1T)
… told Auntie them.

(7) *"Even then,"* Celia told Andrew, *"Sam took some persuading."* (AH1)
"Selv da trengte Sam en god del overtalelse," sa Celia. (AH1T)
… said Celia.

It should be noted that Mukherjee (2005) excludes examples such as (6) and (7) from his classification of ditransitive constructions, limiting cases with direct speech THEMEs to instances like (5) where these follow, rather than precede, or interrupt, the reporting predication (see also Huddleston and Pullum, 2002: 1026). For this contrastive study, I decided to include all three types since one type may be translated by another, as in (7) (see also Bourne, 2002: 245).

Section 4 expands the topic to include TELL predications in French, which resembles both English and Norwegian in containing a TELL verb, *raconter*, and Norwegian, but not English, in containing a SAY verb, *dire*, which occurs in the ditransitive. The data are from the Norwegian–English–German–French part of the Oslo Multilingual Corpus (OMC). This part of the OMC contains a total of 408,558 words from five Norwegian novels, together with their translations into English, French and German (see Johansson, 2007 for details). Two methods are employed to retrieve relevant examples from the OMC. The first method takes as its starting point Norwegian ditransitive *fortelle* constructions, which function as *tertia comparationis* for their English and French translations. The second method starts with all instances of ditransitive *tell* in the English target texts, which are compared to their corresponding French translations and Norwegian sources.

3. English and Norwegian TELL constructions in the ENPC

The original English texts in the ENPC contain 772 examples of *tell*, 449 of which (58%) occur in double object constructions. For Norwegian, the total number for *fortelle* is 536, and 120 examples (22%) of these occur in double object constructions. These types of constructions would therefore seem to be more salient for the English verb. 15% of the English examples and 15% of the Norwegian ones are from non-fiction texts. I chose not to distinguish between fiction and non-fiction texts in this study since the distinction is a crude one and the majority of the relevant examples from non-fiction are found in narrative texts, such as Peter Mayle's *A Year in Provence*.[5] Section 3.1 contains an overview of the various types of THEME and RECIPIENT that occur with these constructions in the source texts in the two languages. Section 3.2 examines the distribution of the two forms of double object construction in the source texts.

[5] One difference between the two sets of texts is the greater tendency for the TELLER in non-fiction to be inanimate, the RECIPIENT generic and the verb in the present tense, yielding a construction that may be paraphrased 'the evidence shows (us) that'.

Section 3.3 presents the translation correspondences of the constructions in the two sets of target texts.

3.1 Types of THEME and RECIPIENT

The RECIPIENT in TELL constructions is almost invariably animate. It is encoded in most cases by either a personal pronoun or a proper noun (93% in English and 87% in Norwegian). The remainder, with a single exception, cited as (8), are either encoded by NPs with an animate head, such as "the boys" and "den gamle kvinnen" (the old woman), or a head metonymically related to an animate, such as "the staff meeting" and "politiet" (the police). (8) is the only example where there is no doubt that the RECIPIENT is incapable of receiving the communication. In this case the TELLER is merely giving voice to her disappointment and frustration.

(8) Jeg forteller *sykkelen* at jeg er gravid, tross Lippes loop. (CL1)
I tell *the bicycle* that I'm pregnant, in spite of the Lippes coil. (CL1T)

As for THEMEs, we have already seen examples in (1)–(7) of the five types of direct object we find with English *tell*: NPs, *that*-clauses, *wh*-clauses, *to*-infinitive clauses, and direct speech. In the original Norwegian texts in the ENPC, there are only three types of THEME, NPs as in (9), *that*-clauses, as in (10), and *wh*-clauses, as in (11).

(9) Og så fortalte hun dem *sin historie*. (TTH1)
Then she told them *her story*. (TTH1T)

(10) Jeg forteller Nick *at jeg skal gifte meg i New York*. (KT1)
I tell Nick *that I am going to get married in New York*. (KT1T)

(11) Han fortalte meg *hvordan det foregikk*. (EG2)
He told me how it happened.
He told me all about *how it's done*. (EG2T)

Fortelle does not occur with a clausal infinitive object to code an instruction. Nor is it used in the original Norwegian texts in a construction with an explicit RECIPIENT to report direct speech. That the ditransitive construction is not impossible in Norwegian is shown by the idiomatic translation in (6). Figure 1 shows the distribution of *tell* and *fortelle* in the various constructions.

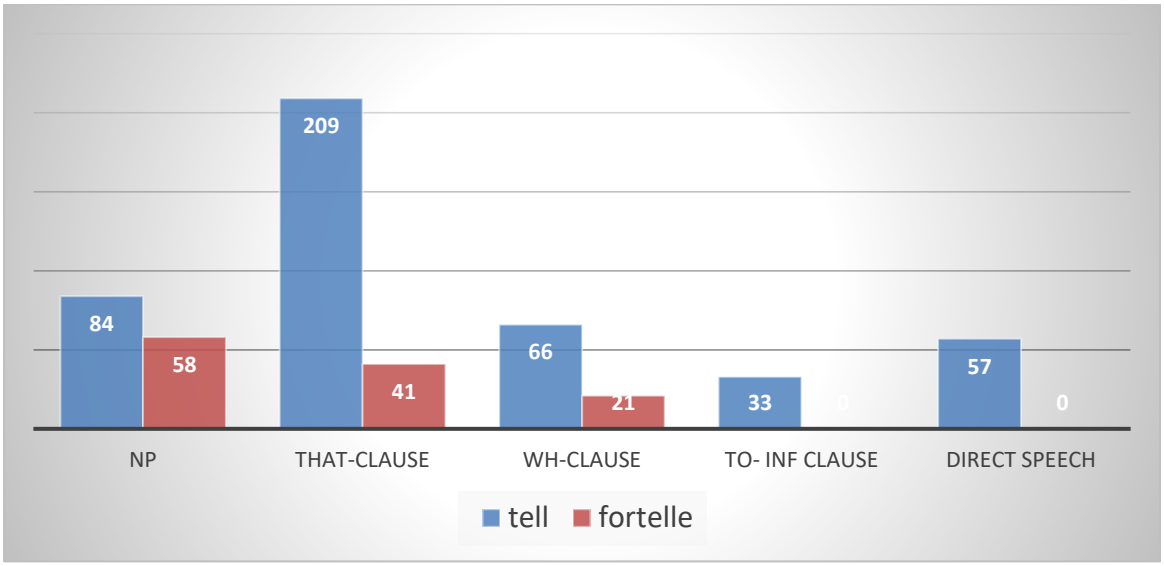

Figure 1. Raw numbers for *tell* (n = 449) and *fortelle* (n = 120) in double object constructions in original texts.

The figure shows that not only is *tell* almost four times as common as *fortelle* in double object constructions, but that the two verbs differ with respect to their relative distribution with the three types of THEME with which both occur. The *that*-clause is the most salient form of direct object of *tell*, occurring in 47% of *tell* double object constructions, and 27% of all occurrences of *tell* (the corresponding proportion in Mukherjee, 2005, based on ICE-GB, is 30%). For Norwegian, the NP form of direct object accounts for 49% of examples of *fortelle* with two objects, but just 11% of all tokens of *fortelle*. The *fortelle* construction with an NP THEME and an explicitly coded RECIPIENT can therefore not be said to represent a particularly salient construction with the verb *fortelle*.

3.2 Ditransitive versus prepositional dative

All examples cited thus far have been of ditransitive constructions. However, both *tell* and *fortelle* can occur with the prepositional dative, as in (12) and (13).

(12) There is *no one* I would ever tell *this to*, except Cordelia. (MA1)
Det er *ingen* jeg ville finne på å *si dette til*, bortsett fra Cordelia. (MA1T)
…say this to…

(13) Men de *fortalte* siden alle detaljer *til alle som ville høre*. (HW2)
…to everyone who wanted to hear.
But they later told the details *to anyone who wanted to know*. (HW2T)

Table 1 shows how often the two verbs occur in the constructions with NP THEMEs (there are no examples in either language of the prepositional dative with a TELL verb and a clausal THEME).

Table 1. Ditransitive vs. Prepositional dative with TELL verbs and NP THEMEs in ENPC.

	Ditransitive		**Prepositional dative**	
tell	83	98.8%	1	1.2%
fortelle	46	79.3%	12	20.7%

Table 1 shows that *fortelle* is much more likely than *tell* to occur in the prepositional dative construction. According to Mukherjee (2005: 123), the prepositional dative is "hardly ever used" in adult speech in English. Indeed, example (12) is the only such example with *tell* among 84 tokens in the ENPC with NP THEMEs. In it the RECIPIENT is the antecedent of a relative clause containing the preposition. In this case the ditransitive ('There is no one I would ever tell this, except Cordelia') would not be felicitous. In eight of the twelve examples of the prepositional dative in Norwegian the THEME is encoded by a pronoun, either *den/det* (it) or *dette* (this). Only two of the RECIPIENTs in these examples are pronominal, and both of these are indefinite *noen* (anyone). Four of the 12 examples receive congruent translations, including three of the four containing full nominal direct objects, as in (14).

(14) Men de *fortalte* siden alle detaljer til alle som ville høre. (HW2)
But they told later all the details to everyone who wanted to hear.
But they later *told* the details to anyone who wanted to know. (HW2T)

While the translation in (14) is perfectly idiomatic, as are the other three prepositional dative translations, the fact that there is only one example of this construction in the original English

texts in the ENPC would suggest that its rarity in spoken English noted by Mukherjee (2005: 123) is also true of the written mode.

3.3 Translation correspondences of *tell* and *fortelle*

The translators of the texts in the ENPC adopt four main strategies in translating double object TELL predications; sometimes they use a syntactically congruent translation containing either the corresponding TELL verb or another verb, and sometimes a syntactically divergent translation containing the corresponding TELL verb or another verb. There are 29 zero translations into Norwegian and two into English. The various strategies will be illustrated in turn, starting in (15)−(16) with congruent translations employing the TELL verb.

(15) Have I *told* you my Theory of Life, by the way? (JB1)
Har jeg *fortalt* deg min Livsteori, forresten? (JB1T)

(16) Jeg *fortalte* deg at vi fant en av dem helt ute i trappehuset. (GS1)
I *told* you we found one of them right on the stairway. (GS1T)

Congruent translations with the TELL verb are used by translators in both directions and with all three types of direct object that occur in both sets of original texts. This form of translation is maximally congruent, with the cognate verb being used in the identical syntactic construction.

In some translations, exemplified here by (17)−(18), the syntax of the original is preserved, but another verb is used instead of *tell/fortelle*. (An overview of alternative verbs is given for English in Table 3 and for Norwegian in Table 4.)

(17) Han *fortalte* det til Henry som lo støyende. (OEL1)
He *said* it to Henry who laughed noisily. (OEL1T)

(18) We wouldn't start *telling* people he was dead until after I'd talked to his lawyers. (DF1)
Vi skulle ikke begynne å *meddele* folk at han var død før jeg hadde snakket med advokatene hans. (DF1T)
…inform…

We saw in section 3.2 that English writers tend to avoid the prepositional dative with *tell*. However, the ditransitive is not an option in (17) because of the relative clause modifying the RECIPIENT.

Sometimes translators prioritise the cognate lexeme at the expense of the construction, by retaining the TELL verb in a syntactically divergent translation, as in (19), in which the THEME is omitted, and (20) which retains the THEME but omits the RECIPIENT.

(19) Det er ikke verdt vi *forteller* henne dette. (THA1)
…tell her this….
We'd better not *tell* her. (THA1T)

(20) I take it your mother *told* you I stopped by. (SG1)
Jeg går ut fra at Deres mor har *fortalt* at eg var innom. (SG1T)
…told that I….

There are no Norwegian translations like (19) that omit the THEME, which is implicit in the co-text. There are, on the other hand, four examples that resemble (20). In these, the translators omit the RECIPIENT, but also insert *om* (about) before the THEME, as shown in (21). There are also seven translations into English and two into Norwegian that retain the TELL verb and the RECIPIENT but encode the THEME in an *om/about* (or *of*) phrase, as in (22)−(23).

(21) "You promised to let Simon *tell* us his problem." (RDA1)
"Du lovte å la Simon *fortelle om sine problemer*." (RDA1T)
…tell about his problems.

(22) Hun *fortalte* det til Sol. (HW1)
…told it to Sol
She *told* Sol *about* it. (HW1T)

(23) He'd *told* me nothing about any love life. (DF1)
Han hadde ikke *fortalt* meg *om* noe kjærlighetsliv. (DF1T)
…had not told me about …

There are five divergent translations into Norwegian that resemble (23) in recoding the THEME in an *om* phrase, but that contain a verb other than *fortelle*, such as *be* (ask) in (24). 59 translations into Norwegian and two into English exhibit the dative alternation (see Table 2), almost always in the direction of the prepositional dative, and normally with a SAY verb, as in (25) and (26). We saw in Table 1 that the prepositional dative is much more common in the original Norwegian texts. This difference in distribution is also reflected in the translated texts, presumably as a result of what Halverson (2017: 14) calls "magnetism", exerted by the structure of the target language.

(24) Mister O'Connell never told us to get out or stay quiet. (RDO1)
Herr O'Connell *ba* oss aldri *om* å komme oss ut eller være stille. (RDO1T)
…asked us never about to get us out…

(25) Slikt kunne jeg ikke *fortelle* mor. (MN1)
…tell mother…
But I couldn't *say* such things to Mother. (MN1T)

(26) I won't pay till you *tell* that boy to apologise to me. (BO1)
Jeg betaler ikke før du *sier* til den gutten at han skal be om unnskyldning. (BO1T)
…say to that boy that he…

There are as many as 131 examples, over a third of the total number of translations into Norwegian, that omit the RECIPIENT and employ an alternative verb, which again is often *si*, as in (27). (28) is one of just eight translations into English which employ this strategy.

(27) Our mother *tells* us which pages to do. (MA1)
Moren vår *sier* hvilke sider vi skal gjøre. (MA1T)
…says which pages…

(28) Søster Vera *fortalte* meg bare at tante var død. (EG1)
Sister Vera told me only that auntie was dead.
Nurse Vera just *said* my aunt was dead. (EG1T)

One final translation strategy that should be exemplified is the greater number of zero translations into Norwegian, 29 (6.5%) as compared to just two (1.9%) into English. Both of the English translations, one of which is cited as (29), contain *ingen* (no one) as the TELLER.

(29) For ingen skulle *fortelle* ham at møblene var fra Ikea eller at nipsgjenstandene kom fra en eller annen basar på Grønland! (EG2)
Because no one was going to tell him that the furniture was from IKEA or the decorations from some second-hand shop or other in Grønland [a district in Oslo].
Because it was clear that neither furnishings nor ornaments came from chain stores. (EG2T)

(30) Andrew *told* his patient, "I'm happy for you, Mary." (AH1)
Andrew snudde seg mot pasienten. "Jeg er så glad, Mary Rowe." (AH1T)
Andrew turned towards the patient. "I am so happy, Mary Rowe".

(31) Will you listen while I *tell* you what is really bothering me? (RDA1)
Vil dere vite hva som virkelig plager meg? (RDA1T)
Do you want to know what's really bothering me?

Example (29) does not describe a genuine act of communication. The expression glossed as 'no one was going to tell him' is idiomatic and indicates a degree of conviction on the part of the RECIPIENT of the falsehood of the predication in the *that*-clause. Both (30) and (31) encode real communicative acts, but in both cases the TELLER and RECIPIENT can be inferred from the context. In other words, these are textbook examples of implicitation (Vinay and Darbelnet, 1995: 344). Pípalová (2012: 83) notes a similar tendency for what she calls the 'reporting frame', by which she means the clause containing a reporting verb such as *tell*, to be omitted with Direct Speech THEMEs in translations from English into Czech and vice versa.

Figures 2 and 3 contain details of the main types of translation strategies employed by the two sets of translators, into English and Norwegian respectively.

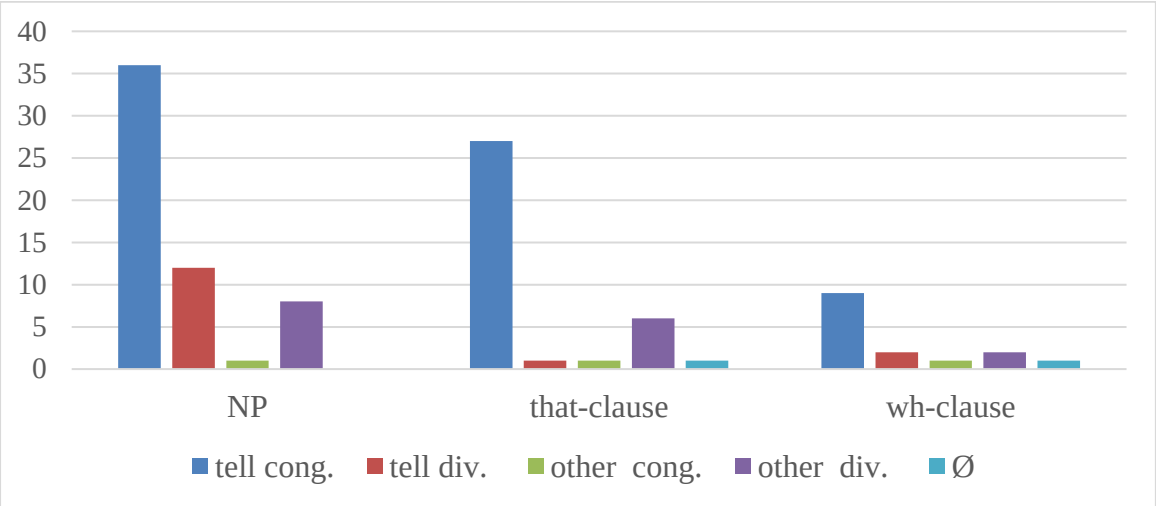

Figure 2. Strategies employed by translators from Norwegian into English in raw numbers.

Figure 2 shows that, for all three double object *fortelle* constructions, the English translators tend to opt for *tell*, most often in a congruent translation. In other words, they faithfully stick both with the construction and with the cognate lexeme. To (over-)generalise, in constructions where the translators can use *fortelle* in Norwegian, they are likely to retain the form of the construction with the verb *tell* in English. When it comes to translations from English into Norwegian, on the other hand, we are confronted in Figure 3 with a much broader palette of options taken by the translators.

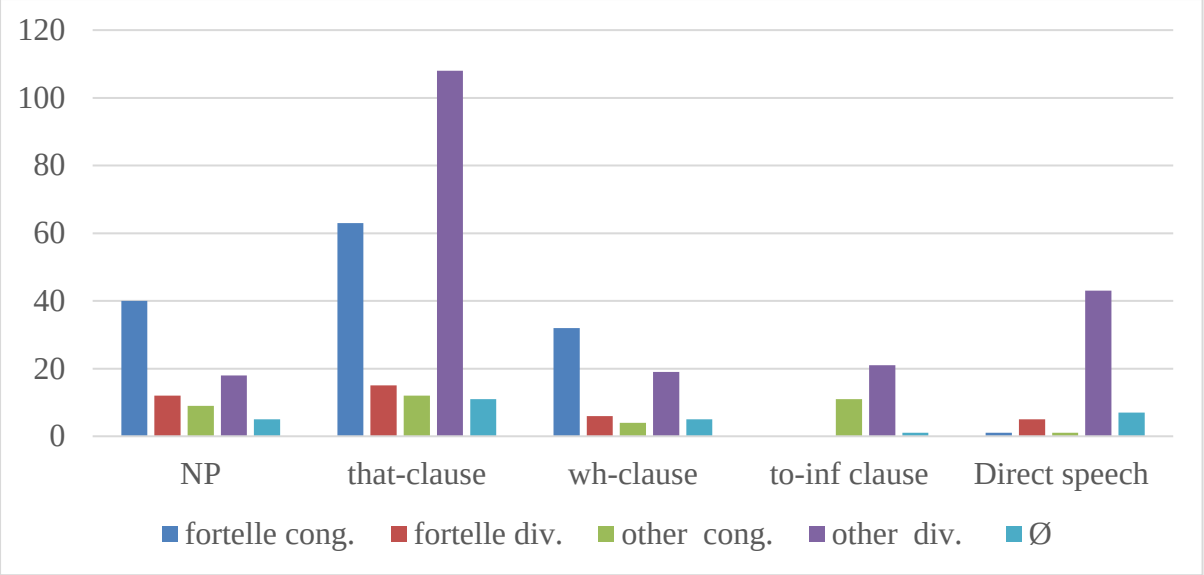

Figure 3. Strategies employed by translators from English into Norwegian in raw numbers.

It is only in the case of NP and *wh*-clausal THEMEs that translators into Norwegian prefer *fortelle* as the target verb. Translations of *that*-clausal THEMEs resemble translations of the two English double object constructions with no Norwegian counterpart in containing more examples with the verb *si* (say). Table 2 contains details of the most common verbs used in four types of divergent translation into both languages.

Table 2. Lexicogrammatical correspondences in four types of divergent translations.

	English translations		**Norwegian translations**	
	Total	**Verbs**	**Total**	**Verbs**
Omission of RECIPIENT	8	*say*: 2 other verb: 6	163	*fortelle* (tell): 32 *si* (say): 109 other verb: 22
Omission of THEME	5	*tell*: 5	0	
Ditransitive to Prepositional Dative	1	other verb: 1	59	*fortelle* (tell): 2 *si* (say): 52 other verb: 5
Prepositional Dative to Ditransitive	1	*tell*: 1	0	

We can see from Figures 3 and 4 and Table 2 that translators in both directions may employ verbs other than the relevant cognate. Tables 3 and 4 contains details of all verbs other than *tell* and *fortelle* which are found in both syntactically congruent and syntactically divergent translations.

Table 3. Verbs other than *tell* used in translations into English.

Number	Verbs
4	*say*
2	*inform*
1 each of	*admit, be reminiscent of, confirm, corroborate, demonstrate, dictate, give, keep secret, object, outline, perceive*
13	Total number of types
17	Total number of tokens

Table 3 shows that there are in all 13 verbs other than *tell* that are used in English translations of *fortelle*. Apart from the general communication verb *say* and the even more general transfer verb *give*, these are, with two exceptions, either troponyms giving more information than *tell* about the mode of communication (*demonstrate*, *outline*) or more information about the attitude of the speaker to the content of the communication (*admit*, *confirm*). The fact that 11 of the 13 verbs in Table 3 are hapaxes indicates that we may here be witness to individual, perhaps even idiosyncratic, choices made by the translators. This impression is reinforced by the fact that two translators are responsible for over half of the verbs other than *tell* and *say* (7 of 12).

Table 4. Verbs other than *fortelle* used in translations into Norwegian.

Number	Verbs
183	*si* (say)
12	*be* (ask)
7	*forklare* (explain)
4	*høre* (hear)
2	*forsikre* (assure), *gi beskjed* (give message), *gjøre klart* (make clear), *svare* (answer)
1 each of	*bekjentgjøre* (announce), *bekrefte* (confirm), *bemerke* (comment), *beordre* (order), *bestemme* (decide), *erklære* (state), *formane* (urge), *få* (get), *innbille* (imagine), *instruere* (instruct) *love* (promise), *lære* (teach), *meddele* (inform), *mene* (think), *minne* (remind), *nevne* (mention), *opplyse* (inform), *overbevise* (persuade), *presisere* (clarify), *proklamere* (proclaim), *påstå* (assert), *replisere* (reply), *servere* (serve), *skrive* (write), *tilstå* (confess), *tro* (believe), *true* (threaten), *utbryte* (exclaim), *vise* (show)
37	Total number of types
242	Total number of tokens

Over half of the translations of *tell* predications into Norwegian contain a verb other than *fortelle*, which may be compared to just 14% of translations into English containing verbs other than *tell* (the difference between the two, according to a chi. sq. test, is significant at the p=0.05 level). The general communication verb *si* (say), which, unlike English *say*, can take an indirect object (Faarlund *et al*., 1997: 726), accounts for 76% of the verbs other than *fortelle*. The

second most common verb *be* (ask) is used to translate English *to*-infinitive instructions as in (24) and (32).

(32) She hardly spoke to him apart from *telling* him to go to the shop for five Woodbines. (ST1)
Hun sa nesten aldri et ord til ham, bortsett fra når hun *ba* ham stikke ned i butikken etter sigaretter til henne. (ST1T)
She said hardly ever a word to him, apart from when she asked him to pop down to the shop for cigarettes for her.

The choice of an ASK verb instead of a TELL verb would appear to alter the illocutionary force of the THEME from an instruction to a request. Since the 12 examples are spread over 11 translators, this change cannot be ascribed to the idiolects of a handful of translators. The most likely explanation for this behaviour is that the ASK verbs in the two languages differ in their force, but it lies outside the scope of this paper to pursue this putative distinction.

There are four examples in which the RECIPIENT is recoded from indirect object of *tell* to subject of *hear*. This is an example of the translation technique called 'modulation' by Vinay and Darbelnet (1995: 346), whereby a participant in the source text is encoded in a translation in a different semantic role (as in 'She told it to him' –> 'He heard it from her'). The remainder of the verbs in Table 4 resemble the troponyms used in the translations into English in Table 3 by either giving more information than *fortelle* about the mode of communication (*utbryte* (exclaim), *skrive* (write)) or more information about the attitude of the speaker to the content of the communication (*formane* (urge), *tilstå* (confess)). The total of 37 different types chosen to translate 449 tokens may be compared to the 47 types employed by translators into English of 382 Dutch tokens of *beginnen* (begin/start), another common lexeme, in Vandevoorde's study of inchoativity (2020: 82). Vandevoorde actually decided to discard hapaxes from her study of *beginnen*, regarding these as expressions of translators' idiosyncracies. The 29 hapaxes in my study are dispersed over 16 texts, one translator being responsible for six of them and a further two for three each, indicating that some translators are more idiosyncratic than others.

Table 5 shows the degree of overall mutual lexical and syntactic correspondence (see Altenberg, 1999; Ebeling and Ebeling, 2013: 27) between *tell* and *fortelle* in double object constructions, as well as the correspondence in each of the three constructions that occur with both verbs in the ENPC.

Table 5. Mutual correspondence of *tell* and *fortelle*.

	Lexical mutual correspondence	**Syntactically congruent translations**	**Lexical + syntactic mutual correspondence**
In examples with NP THEMEs	71.1%	61.3%	54.2%
In examples with *wh*-clausal THEMEs	62.8%	59.3%	53.5%
In examples with *that*-clausal THEMEs	44.4%	43.2%	38%
Overall in double object constructions	56%	51%	39%

Table 5 shows that there is a relatively high degree of mutual lexical correspondence between the two verbs in constructions with NP THEMEs. This is also the type of THEME that gives rise

to the highest incidence of syntactically congruent translations. However, when comparing the percentages in the fourth column to the second and third, we see that the choice of the congruent construction appears more likely to prompt the use of the cognate verb rather than vice versa (the difference between the two is syntactically significant, according to a chi.sq. test at the level of p=0.05). With respect to both types of clausal THEMEs, there is a lower degree of both lexical and syntactic correspondence than is the case for NP THEMEs. However, in both cases the two types of correspondence seem to go hand in hand (in both cases the probability of their being a difference between them is >0.1). That is, if translators of predications containing clausal THEMEs first opt for the cognate verb or the corresponding syntactic construction, they are likely to combine the two in their translation.

4. French correspondences of Norwegian originals and English translations

This section is divided into two parts. Section 4.1 contrasts English and French translations of Norwegian original examples of double object *fortelle* constructions. Section 4.2 takes as it starting point double object *tell* constructions in the English translations and compares these to the corresponding Norwegian originals and the French translations of these. (There are no English or French source texts in the Norwegian–English–German–French part of the OMC). It should be mentioned at the outset that shorter extracts from three of the texts in the OMC are included in the ENPC, so any contrasts between English and Norwegian that emerge from the data cannot be viewed as independent of those described in section 3.

4.1 English and French translations of Norwegian *fortelle* predications

There are 94 examples of double object *fortelle* constructions in the original Norwegian texts in the OMC, 87 of which are ditransitive, the remaining seven being the prepositional dative. Table 6 contains an overview of the English and French verbs used in the two sets of translations.

Table 6. Verbs used in congruent and divergent English and French translations of double-object *fortelle* predications.

English		French	
83	*tell* (76 Cong., 7 Div.)	39	*dire* (say: 32 Cong., 7 Div.)
2	*let know* (Cong.)	28	*raconter* (tell: 25 Cong., 3 Div.)
1	Cong.: *give, inform, say* Div.: *emphasise, indicate, narrate, point out, talk*	5	*expliquer* (explain: 2 Cong., 3 Div.)
1	Ø	4	*révéler* (reveal: Cong.)
		3	*parler* (speak: 2 Cong., 1 Div.)
		2	*faire comprendre* (give to understand: Cong.)
		1	Cong.: *affirmer* (affirm), *faire savoir* (give to understand), *laisser à penser* (give to think), *ne cacher* (not conceal), *répéter* (repeat) Div.: *addresser* (address), *indiquer* (indicate), *pouvoir savoir* (let understand), *prévenir* (warn)
		4	Ø
94	Cong. 81, Div. 12, Ø 1	94	Cong. 72, Div. 18, Ø 4

According to the data in Table 6, both sets of translators employ congruent constructions, as in (33), in the majority of cases. The difference between the behaviour of the translators in this respect is marginally significant at the 0.05 level (p=0.046 according to a chi.sq. test). While English translators are more likely to employ a congruent translation of the ditransitive, the opposite is the case with the prepositional dative, with French translators employing congruent translations in five of the seven, and English translators in just two. Four of the divergent English translations contain the ditransitive, as in (34).

(33) *Fortell* meg straks hvor han er. (NF1)
Tell me at once where he is.
Tell me where he is. (NF1TE)[6]
Dites-moi vite où il est. (NF1TF)
Say me quickly where he is.

(34) Jeg trengte én å *fortelle* det til at Ana er død. (JG3)
I needed someone to tell it to that Ana is dead.
I needed to *tell* someone that Ana is dead. (JG3TE)
J'avais besoin de *dire* à quelqu'un qu'Ana était morte. (JG3TF)
I had the need to say to someone that Ana was dead.

Among the French translations of the prepositional dative there is one, (35), which contains a double coding of the RECIPIENT. This sort of double coding does not occur in either English or Norwegian, but can be used in French when the speaker wishes to emphasise the RECIPIENT.

(35) Det *fortalte* han til deg? (JG3)
That told he to you?
He *said* that to you? (JG3TE)
Il t'a *raconté* ça, à toi? (JG3TF)
He you told that, to you?

The results in Table 6 are in line with those in section 3.3, which showed that *tell* is by far the most likely English translation of *fortelle*. In fact, (35) is the only case where the English translator employs the verb *say*. In French, on the other hand, the verb *dire* is more commonly used than the verb *raconter*. The difference between the two languages with respect to these verbs of SAYING and TELLING is illustrated in Table 7.

Table 7. Basic English and French verbs of TELLING and SAYING.

Verbs of TELLING / SAYING	**English target texts**	**French target texts**
+ RECIPIENT	*tell*	*raconter*
		dire
– RECIPIENT	*say*	

The categorisation in Table 7, although it employs the notation of semantic componential analysis, is merely intended to illustrate the prototypical senses of the communication verbs. The verb that displays the least constructional variation in the translated data is *raconter*, which only occurs in three divergent translations, according to Table 6. In all three of these the prepositional dative replaces the ditransitive, as in (36).

[6] 'TE' and 'TF' stand for translated text in English and French respectively.

(36) Lorch *fortalte* Dina mange rare ting. (HW2)
Lorch *told* Dina many strange things. (HW2T)
Lorch *racontait* à Dina des tas de choses bizarres. (HW2TF)
Lorch told to Dina lots of strange things.

Table 8 contains details of how often the two French verbs *dire* and *raconter* are used to translate the three syntactic types of THEME in the Norwegian originals.

Table 8. Types of THEME in Norwegian original texts translated by *dire* and *raconter*.

	NP	***that*-clause**	***wh*-clause**	**Total**
dire	8	12	19	39
raconter	20	3	5	28

We see in Table 8 that *raconter* is more than twice as likely as *dire* to occur with an NP THEME. Moreover, eight of these 20 NPs contain the word *historie* (story), either standing alone, as in (37), or as part of a compound, such as *godnatthistorie* (bedtime story) or *løgnhistorie* (fake story).

(37) Så hadde han bestemt seg og ga seg til å *fortelle* meg en historie. (BBH1)
Then he had made up his mind and set about telling me a story.
Then his decision was made, and he *told* me a story. (BBH1TE)
Puis, ayant pris sa décision, il s'est mis à me *raconter* une histoire. (BBH1TF)
Then, having made his decision, he started me telling a story.

The fact that the various events in a story are told consecutively serves to explain the common origin of the present-day French verbs *(ra)conter* and *compter* (to count), as it does the relationship between the Norwegian verbs *fortelle* and *telle* (to count).[7] As for English, according to the OED the verb *tell* was used in the sense 'to count', side by side with the recount sense, from Old English up until the eighteenth century.[8]

4.2 English translations, Norwegian originals and their French translations

According to Table 6 there are 83 examples of Norwegian double object *fortelle* predications translated into English by *tell*. Three translations substitute the ditransitive for the prepositional dative, the remaining four contain just one object. This leaves 79 double object *tell* examples translating double object *fortelle*. In actual fact, however, these 79 examples represent fewer than a third of the 272 examples of double object *tell* constructions in the translations in the OMC. In this section, we will examine what sort of constructions trigger the other 193 uses of *tell*, and how the French translators respond to these same verbal triggers. Five of the examples translate monotransitive *fortelle*, leaving 188 examples containing other verbs. The verb *si* (say) occurs in 102 (37.5%) of the Norwegian originals. Table 9 contains details of three *si* constructions, monotransitive (MT), ditransitive (DT) and prepositional dative (PD), and their French translations.

[7] The nouns *conte* (French), *fortelling* (Norwegian) and *tale* (English) are all related to the TELL verbs and all denote informal stories, often oral in origin.
[8] The expression 'to tell the time' displays a fossilised use of the 'count' sense. It originally referred to the practice of ascertaining the time by counting the ringing of the church bells.

Table 9. French translations of Norwegian *si* (say) predications translated into English by double object *tell*.

Norwegian originals		French translations	
51	MT *si*	19	DT *dire* (say)
		12	MT *dire* (say)
		1 each of	MT: *ajouter* (add), *éclairer* (throw light on), *expliquer* (explain), *faire valoir* (maintain), *raconter* (tell), *répeter* (repeat), *répondre* (answer) DT: *annonser* (announce), *avouer* (swear), *demander* (ask), *donner* (give), *expliquer* (explain), *ordonner* (order), *prévenir* (warn) PD: *ordonner* (order) IT*: *demander* (ask)
		4	Ø
27	DT *si*	19	DT *dire* (say)
		1 each of	MT: *répondre* (answer), *constater* (remark) DT: *montrer* (show), *écriver* (write) PD: *asséner* (strike)
		3	Ø
24	PD *si*	9	DT *dire* (say)
		3	PD *dire* (say)
		1	DT + PD *dire* (say)
		1	MT *dire* (say)
		2	DT *répéter* (repeat)
		1 each of	MT: *donner l'ordre* (order), *penser* (think), *souffler* (murmur) DT: *avouer* (swear), *convaincre* (convince), *rassurer* (reassure)
		2	Ø

* intransitive

63% of the translations of *si* in Table 9 contain the verb *dire*, with the ditransitive *dire* construction being the most frequent translation, irrespective of whether the Norwegian originals are monotransitive, as in (38), ditransitive, as in (39), or prepositional dative, as in (40).

(38) — Hva var det jeg *sa*? utbrøt hun. (JG3)
— *What was it I said? she burst out.*
"What did I *tell* you?" she cried. (JG3TE)
— Qu'est-ce que je t'avais *dit*! s'exclama-t-elle. (JG3T)
What is it I you said! exclaimed she.

(39) Men øynene hennes *sa* ham hvem han tilhørte. (HW2)
But her eyes said him who he belonged to.
But her eyes *told* him to whom he belonged. (HW2TE)
Mais ses yeux lui *disaient* à qui il appartenait. (HW2TF)
But her eyes him said to whom he belonged.

(40) Jeg *sa* til meg selv at jeg måtte glemme denne dagen. (NF1)
I said to myself that I must forget this day.
I *told* myself I would have to forget that day. (NF1TE)
Je me *dis* que je devais oublier cette journée. (NF1TF)
I me say that I should forget this day.

In both (38) and (40), the salient correspondence of *si* in French, *dire*, is utilised, but not the most similar syntactic construction. Most striking perhaps is the extent to which the translators

resort to a ditransitive construction to translate monotransitive *si*.[9] There would appear to be a felt need to mention the RECIPIENT explicitly on the part of translators into English and French, both sets of whom appear to be influenced by the "gravitational pull" of the target language grammar (Halverson 2007), rechristened "magnetism" by Halverson (2017).

Before looking at other Norwegian forms that are translated into English by double object *tell*, I should point out that the *si* originals in Table 9 represent just 5% of the total number of occurrences containing the verb *si* in the original Norwegian texts. The majority of these are monotransitive and are translated into English by monotransitive *say*, and into French by monotransitive *dire*. In all over 20 French verbs are used to translate the Norwegian *si* predications in Table 9, all of which correspond to English double object *tell* translations. All of these verbs are more explicit with respect to the mode or force of communication than Norwegian *si*. According to Nádvorníková (2020) the proportion of neutral reporting verbs (like *say* and *tell*) in original texts in her corpus was 60% for English and 50% for French, indicating a preference for a wider variety of reporting verbs in French, a preference she found reflected in a greater degree of explicitation in translations from English to French than vice versa (Nádvorníková, 2020: 223).

In addition to *fortelle* and *si*, there are in all 36 Norwegian verb types that give rise to double object *tell* translations. These account for 79 tokens. There are six zero translations and one where the English verbal predication translates a Norwegian nominal. Considerations of space dictate that just the two most common of these Norwegian verbs will be exemplified here. These are *forklare* (explain) with 14 tokens, and *be* (ask) with ten tokens. Of the 14 examples of *forklare*, 13 are translated into French by *expliquer*, as in (41). *Expliquer* is also occasionally used to translate *fortelle* (Table 6) and *si* (Table 9).

(41) Jeg *forklarte* Idun at hun måtte skynde seg. (BHH1)
I explained Idun that she had to hurry herself.
I *told* Idun to hurry up. (BHH1TE)
J'ai *expliqué* à Idun qu'elle devrait faire vite. (BHH1TF)
I explained to Idun that she had to make haste.

While *forklare* is almost always translated by French *expliquer* (this is also the case in the 75 examples where it does not prompt a *tell* construction in English), there is no single French verb that stands out in translations of Norwegian *be* (ask). Of the ten examples, three are translated by *prier* (beg), as in (42), and two by *demander* (ask), as in (43).

(42) Jeg *ber* ham ta det rolig. (NF1)
I ask him to take it easy.
I *told* him to compose himself. (NF1TE)
Je le *prie* de rester tranquille. (NF1TF)
I him beg to stay relaxed.

(43) Til sist *bad* jeg henne om å holde opp. (JG3)
Finally asked I her to cut it out.
Finally I *told* her to stop. (JG3TE)
Je lui ai *demandé* d'arrêter. (JG3TF)
I her (have) asked to stop.

The French translations of (42) and (43) incorporating *prier* and *demander* appear to mirror the semantics of the source texts. The English translations appear to alter, and not merely explicitate, the semantics of the reporting verb, as discussed in relation to the opposite direction

[9] This preponderance of the ditransitive is in line with the findings of Malvar Mouco and Pino Serrano (2006: 560). 16.5% of *dire* tokens in their material are ditransitive, compared to 1.7% prepositional dative.

of translation (of *tell* by *be*) in section 3.3. (See Winters, 2007: 420 for a discussion of some translations between German and English that alter the semantics of verbal predications.)

There are six cases where English *tell* does not translate directly a predication in the Norwegian original. In some of these, as in (44), the French translation is faithful to the original, in others, such as (45), there is no expression in either English or French corresponding to the Norwegian expression.

(44) Eller at hun *satte* dem *i gang med* å telle alle ting som var i rommet. (HW2)
Or that she got them going with counting all the things that were in the room.
Or she *told* them to start counting everything in the room. (HW2TE)
Ou encore elle les *mettait* à compter tous les objets qui se trouvaient dans la pièce. (HW2TF)
Or else she them put to count all the objects that were located in the room.

(45) *Erfaringsmessig* er det naturligvis umulig at Ana var den gamle malerens modell. (JG3)
Experience-wise is it naturally impossible that Ana was the old painter's model.
Experience *tells* us it is inconceivable that Ana was the Old Master's model. (JG3TE)
Naturellement, nous *savons* par expérience qu'il est impossible qu'Ana ait été le modèle du peintre. (JG3TF)
Naturally, we know from experience that it is impossible that Ana has been the model of the painter.

The translators of (45) have chosen different options, both of which were available to both of them: that is the English translator could have written "we know from experience that" and the French one "l'expérience nous dit que".

To summarise this section on the Norwegian originals underlying English *tell* translations and their corresponding French translations, two points stand out. The first is the large number of verbs in both the Norwegian and French versions that correspond to English double object *tell* predications. The second is the number of SAY verbs (*si* and *dire*) that correspond to *tell*. In fact, if one were to subject *si* and *dire* to an Anglo-centric classification, it might be more accurate to label them TELL verbs, as they both occur in the ditransitive construction as well as the prepositional dative. Of the two, *dire* is the more common in my data. Thus, to the somewhat impressionistic picture of the relationship between English and French verbs of SAYING and TELLING in Table 7, we can now add information about their Norwegian correspondences, as in Table 10.

Table 10. Basic English, Norwegian and French verbs of TELLING and SAYING.

Verbs of telling/saying	English	Norwegian	French
+ **RECIPIENT**	*tell*	*fortelle*	*raconter*
			dire
		si	
– **RECIPIENT**	*say*		

5. Summary and conclusion

This paper has presented the results of two related studies of ditransitive communication constructions. The first study, presented in section 3, is based on data from the ENPC and contrasts the occurrences of double object constructions containing the cognate verbs English *tell* and Norwegian *fortelle*, both in original texts in the two languages and in translations of these. The second study, presented in section 4, is based on data from the OMC and contrasts English and French translations of Norwegian double object *fortelle* predications. It also looks at all the double object *tell* constructions in the English translations and compares these to the corresponding Norwegian originals and the French translations of these.

Three research questions were presented in section 1. The first question asked about the degree of similarity between the verbs *tell* and *fortelle* with two objects in the original texts in English and Norwegian. It turns out that there are considerable differences in the lexicogrammatical behaviour of the two verbs in double object constructions. They differ much more in their lexicogrammar than the two GIVE lexemes *give* and *gi*. *Tell* is more than four times as common and occurs with a greater syntactic variety of THEMEs than *fortelle*, while the latter occurs much more often than *tell* in the prepositional dative construction.

The second question asks whether there are some kinds of tokens that are usually translated by congruent constructions. It was answered in section 3.3 by comparing the target texts in Norwegian and English with their sources. It transpires that tokens with NP THEMEs are those most often translated congruently, 65% in the direction Norwegian–>English and 60% in the direction English–>Norwegian. This is also the type of THEME that sees the most translations containing the cognate lexeme. Moreover, a decision on the part of translators to employ the congruent construction increases the likelihood of their employing the cognate verb. Perhaps the most striking feature of the Norwegian translations is the tendency to employ the verb *si* (say). This tendency is no doubt facilitated by the fact that *si*, unlike its English cognate *say*, can take an indirect object, enabling translators to retain the syntactic construction in their translations, while opting for a more common neutral reporting verb. Translators into Norwegian also employ a greater variety of verbs to translate *tell* than do English translators of *fortelle* predications.

The third question asked about the French translation correspondences of the English and Norwegian constructions. It was answered in section 4 by comparing the French and English translations to one another and to their Norwegian sources. The results show that French resembles Norwegian in several respects. In the first place since the verb *dire*, like Norwegian *si*, can take an indirect object, this renders it an appropriate correspondent of many English ditransitive *tell* predications. In the second place French contains a more specialised TELL verb, *raconter*, which resembles Norwegian *fortelle* in being more restricted in its distribution than English *tell*. And thirdly, there is a large number of other verbs that are used to translate the Norwegian originals that give rise to double object *tell*. There is, however, one respect in which the French translations resemble their English counterparts, namely the addition of an explicit RECIPIENT in translations of original Norwegian monotransitive sentences.

To round off, it is appropriate to ask what, if anything, this study has contributed to our knowledge of double object communication constructions. The fact that English *tell* corresponds in large measure to Norwegian *si* and French *dire* comes as no surprise. Indeed, *dire* precedes *raconter* in the French definition of *tell* in the bilingual *Concise Oxford-Hachette French Dictionary* (1998). More surprising, given the evidence of Table 6 that *dire* is the single most common verb used to translate double object *fortelle*, is the omission of *dire* from the definition of *fortelle* in the bilingual Norwegian–French dictionary *Fransk Blå Ordbok* (2002), which gives *raconter* and *faire le récit de* as its primary correspondents. Another point worth noting is the number of Norwegian and French communication verbs that correspond to English

tell, in addition to the large number of correspondences with *si* and *dire*, which serves to further underline the polysemous nature of the English verb.

In order to get a fuller picture of the distribution of the cognate verbs *tell* and *fortelle*, and of French *raconter*, future studies should take into account their occurrence in monotransitive and passive constructions. It would also be an advantage to expand the data to include SAY verbs in all three languages as well as original texts in all three, since the sub-corpus of the OMC used in the present study is mono-source. A final point concerns the broader topic of cognate verbs in English and Norwegian that partake of the dative alternation. It was shown in Egan (forthcoming) that the distribution of the physical transfer GIVE verbs was very similar in the two languages. We have now seen that the distribution of the message transfer TELL verbs is quite different. Further studies will flesh out the picture by investigating verbs such as those of temporary transfer, LEND verbs, those of ownership transfer, SELL verbs, and those of accompanied transfer, BRING verbs.

References

Åfarli, T.A. 1992. *The Syntax of Norwegian Passive Constructions*. Amsterdam: John Benjamins.

Altenberg, B. 1999. Adverbial Connectors in English and Swedish: Semantic and Lexical Correspondences. In *Out of Corpora: Studies in Honour of Stig Johansson*, H. Hasselgård and S. Oksefjell (eds), 249–268. Amsterdam: Rodopi.

Andersen, M., Fikkert, P., Mykhaylyk, R. and Rodina, Y. 2012. The Dative Alternation in Norwegian Child Language. *Nordlyd* 39:1: 24–43.

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

Bourne, J. 2002. He said, she said: Controlling Illocutionary Force in the Translation of Literary Dialogue. *Targets* 14:2: 241–261.

Bresnan, J. and Hay, J. 2008. Gradient Grammar: An Effect of Animacy on the Syntax of *Give* in New Zealand and American English. *Lingua* 118: 245–259.

Bresnan, J. and Ford, M. 2010. Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language* 86:1: 168–213.

Brøseth, H. 1998. Dobbelt objekt og tilgrensende konstruksjoner i moderne norsk. In *MONS 7: Utvalde artiklar frå det 7. møtet om norsk språk i Trondheim 1007*, J.T. Faarlund, B. Mæhlum and T. Nordgård (eds), 13–34. Oslo: Novus.

Ebeling, J. 1998. Using Translations to Explore Construction Meaning in English and Norwegian. In *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*, S. Johansson and S. Oksefjell (eds), 169–195. Amsterdam: Rodopi.

Ebeling, J. and Ebeling, S.O. 2013. *Patterns in Contrast*. Amsterdam: John Benjamins.

Egan, T. forthcoming. Giving in English and Norwegian: A Contrastive Perspective. In *Ditransitive Constructions in Germanic Languages*, M. Röthlisberger, E. Zehentner and T. Colleman (eds), Amsterdam: John Benjamins.

Faarlund, J.T., Lie, S. and Vannebo, K.I. 1997. *Norsk Referansegrammatikk*. Oslo: Universitetsforlaget.

Halverson, S.L. 2007. Investigating Gravitational Pull in Translation: The case of the English Progressive Construction. In *Text, Processes, and Corpora: Research Inspired by Sonja Tirkkonen-Condit*, R. Jääskeläinen, T. Puurtinen and H. Stotesbury (eds), 175–196. Joensuu: University of Eastern Finland.

Halverson, S. L. 2017. Gravitational Pull in Translation. Testing a Revised Model. In *Empirical Translation Studies. New Methodological and Theoretical Tradition*, G. De Sutter, M-A. Lefer and I. Delaere (eds), 9–46. Berlin: Mouton de Gruyter.

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

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

Levin, B. 1993: *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
Lohndal, T. 2011. Freezing Effects and Objects. *Journal of Linguistics*. 47:1: 163–199, 249–250.
Malvar Mouco, S. and Pino Serrano, L. 2006. Dire et répondre, un couple à comparer. In *Studies in Contrastive Linguistics: Proceedings of the 4th International Contrastive Linguistics Conference: Santiago De Compostela, September, 2005*, C. Mourón Figueroa and T. Iciar Moralejo Gárate (eds), 557–566. Santiago de Compostela: Universidade de Santiago de Compostela.
Mayle, P. 1989. *A Year in Provence*. London: Hamish Hamilton.
Mukherjee, J. 2005. *English Ditransitive Verbs: Aspects of Theory, Description and a Usage-based Model*. Amsterdam: Rodopi.
Nádvorníková, O. 2020. Differences in the Lexical Variation of Reporting Verbs in French, English and Czech Fiction and their Impact on Translation. *Languages in Contrast*, 20: 2: 209–234.
Paradis, M. 2004. *A Neurolinguistic Theory of Bilingualism.* Amsterdam: John Benjamins.
Pípalová, R. 2012. Framing Direct Speech: Reporting Clauses in a Contrastive Study. *Prague Journal of English Studies* 1:1: 75−107.
Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Röthlisberger, M., Grafmiller, J. and Szmrecsanyi, B. 2017. Cognitive Indigenization Effects in the English Dative Alternation. *Cognitive Linguistics* 28:4: 673–710.
Szmrecsanyi, B., Grafmiller, J, Bresnan, J., Rosenbach, A., Tagliamonte, S. and Todd, S. 2017. Spoken Syntax in a Comparative Perspective: The Dative and Genitive Alternation in Varieties of English. *Glossa* 2:1: 1–27.
Tungseth, M.E. 2008. *Verbal Prepositions and Argument Structure: Path, Place and Possession in Norwegian.* Amsterdam: John Benjamins.
Vandevoorde, L. 2020. *Semantic Differences in Translation: Exploring the Field of Inchoativity*. Berlin: Language Science Press.
Viberg, Å. 1996. Cross-linguistic Lexicology. The Case of English *go* and Swedish *gå*. In *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4–5 March 1994*, K. Aijmer, B. Altenberg and M. Johansson (eds), 151–182. Lund: Lund University Press.
Vinay, J-P. and Darbelnet, J. 1995. *Comparative Stylistics of French and English: A Methodology for Translation.* Translated and edited by J.C. Sager and M-J. Hamel. Amsterdam: John Benjamins.
Winters, M. 2007. F. Scott Fitzgerald's Die Schönen und Verdammten: A Corpus-based Study of Speech-act Report Verbs as a Feature of Translators' Style. *Meta: Journal des traducteurs* 52: 412−425.

Corpora and dictionaries

English-Norwegian Parallel Corpus
https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/ [Last accessed 1 June 2021].
Oslo Multilingual Corpus
http://www.hf.uio.no/ilos/english/services/omc/ [Last accessed 1 June 2021].
Dictionnaire de l'Académie, neuvième édition. http://atilf.atilf.fr/academie9.htm [Last accessed 1 June 2021].
Corréard, M.-H. and Grundy, V. 1998. *The Concise Oxford-Hachette French Dictionary*. Oxford: Oxford University Press.
Elligers, A. and Jacobsen, T. 2002. *Fransk Blå Ordbok*. Oslo: Kunnskapsforlaget.

Author's address

Thomas Egan
Department of Humanities
Inland Norway University of Applied Sciences
Holsetgata 31
NO-2318 Hamar
Norway
thomas.egan@inn.no

Moving up and down in real space

The verbal particles *upp* ‘up’ and *ner* ‘down’ and the typological profile of Swedish

Åke Viberg

Uppsala University (Sweden)

The paper focuses on the role of the Swedish spatial particles *upp* ‘up’ and *ner* ‘down’ to signal the endpoint-of-motion in the description of motion situations and is based on Swedish original fiction texts and their translations into English, German, French and Finnish. Frequently the endpoint is marked with a locative preposition such as *på* ‘on’ or *i* ‘in’, and then a particle is required to signal change-of-place. In German and Finnish, the particle is often zero translated and change-of-place is indicated by case. The particle is often zero translated also in French, a V(erb)-framed language. This leads to contrasts at the conceptual level since verticality is not expressed. The result points to radical intra-typological differences between S(atellite)-framed languages in the expression of Path depending on general morpho-syntactic differences. Another important conclusion is that several different classes of motion verbs must be distinguished even in S-languages to describe the expression of change-of-place.

Keywords: multilingual contrastive study, lexical semantics, spatial particles, vertical movement, English/Finnish/French/German/Swedish

1. Introduction

The typological profile of a language is a kind of multilingual comparison that accounts for the distinctive character of the structure of a language in relation to other languages based on work in general typology and contrastive studies (Viberg, 2006, 2013a). The approach can be compared to Hawkins (1986) comparative typology and to the combinations of contrastive studies with other types of language comparison discussed in König (2012) and van der Auwera (2012).

This paper will account for the use of the Swedish spatial verbal particles *upp* ‘up’ and *ner* ‘down’ from this perspective and is part of a general study of such particles in Swedish (see Viberg, 2015a, 2017). Spatial verbal particles represent a prominent characteristic of English, German and Swedish, which will be contrastively compared in this study. To give perspective, data from French and Finnish are included but with a focus on correspondence at

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.v11i1.3439.

the conceptual level, since direct structural counterparts are missing in French and have an unclear grammatical status and relatively low frequency in Finnish.

1.1 Earlier studies

Verbal particles have been studied from many perspectives (see Cappelle, 2005 and Luo, 2019 for English). Particle placement has been a central topic in English (Gries, 2003) but also in the other Germanic languages including the alternation between free and bound forms (Dehé, 2015). In Swedish, the alternation between free forms (e.g. *komma upp* 'come up') and bound forms (e.g. *uppkomma* 'up-' + 'come', 'come into existence') was discussed as part of the general study of particles in Viberg (2017). Within a separate tradition, the central topic has been the semantic networks of the extended meanings of spatial particles since Lindner's (1981) seminal study of *out* and *up*. Swedish particles are studied in depth from the same perspective in a monograph by Strzelecka (2003). A third tradition in which particles play an important role is represented by the studies of the motion situation inspired by Talmy (1985). Particles figure in such studies as one of the realizations of path, although the question of their exact function has not always been addressed. This article will focus on the functions of *upp* and *ner* in descriptions of the motion situation.

Motion has been one of the most frequently studied semantic fields within lexical (or semantic) typology following Talmy (1985, 2000) and Slobin (1996, 2004), see Ibarretxe-Antuñano (ed. 2017) and Matsumoto and Kawachi (eds 2020) for a number of recent studies. Talmy's model of the motion situation can briefly be summed up as follows: The Figure is the moving entity or the entity which has a certain location. (*Peter* in *Peter ran home* and *the book* in *The book is on the table*). Ground refers to the Source, Goal, or Location with respect to which something moves or is located, and Path to the spatial relation between the Figure and Ground as this is expressed through satellites of the verb such as particles and prepositions or incorporated into verbs such as *enter* (IN) and *descend* (DOWN). The expression of Path forms the basis for the typological distinction between verb-framed languages (V-languages), where path of motion is encoded in the verb root (e.g. French *monter* 'move=up'), and satellite-framed languages (S-languages), where the path of motion is expressed in satellites outside the verb, such as particles (e.g. English *go up*). There is also a third type, equipollently-framed languages (E-languages), where Manner and Path are expressed in serial verb constructions.

The exact definition of 'satellite' has been discussed a great deal, but today any element outside the verb root is often regarded as a satellite, and this is the alternative that will be followed in this study. Another disputed point is the very existence of specific types, since the division forms a continuum. It is now generally agreed that most languages have constructions of several types. However, Talmy originally claimed that the classification should be based on the dominant type of construction in a language and it remains an open question to what extent a dominant structure can be identified (see Slobin, 2017 for a recent review of various positions on these issues). The motion situation has also been approached from another direction by Beavers *et al*., 2009, who stress the importance of the motion-independent morphological, lexical and syntactic resources made available by individual languages to express spatial notions.

From a typological perspective, the motion situation in Swedish has been studied in Blomberg (2014), Fagard *et al*. (2013) and Viberg (1992, 2013a). Olofsson (2018) is a construction-grammar study of the motion situation based on large Swedish corpora (see also Teleman *et al*., 1999: 974–979.) The present study is a continuation of a corpus-based study in Viberg (2015a). Boers (1996) is a general study of the semantics of *up* and *down* in English. The motion situation in German is analyzed in Fagard and Kopecka (2021), and the situation in French is discussed from several new perspectives in Aurnague and Stosic (eds 2019) (see

also Lebas and Cadiot, 2003 on *monter* 'move=up'). Finnish has words with a similar function as the Swedish particles *upp* and *ner*, but their grammatical status is unclear (Kolehmainen, 2005) and they have relatively low frequency. Finnish is interesting in particular because of its spatial cases and case-marked adpositions (see Huumo and Ojutkangas, 2006), which functionally fulfill some of the functions of particles in Swedish.

In the description of the Path, the spatial particles interact with other types of spatial markers that express other types of spatial concepts. Levinson and Wilkins (2006) identify three frames of reference. The intrinsic frame describes the relation between a Figure and a Ground, as for example the prepositions *on* and *in* (*the apple in the bowl*). The two other frames introduce an external reference point, which can be relative or absolute. The relative frame often refers to a viewer as in *the stone in front of the tree* (stone on the same side of the tree as the viewer) and *the stone behind the tree* (stone on the opposite side). The absolute frame refers to constant reference points, such as the cardinal points. Another example is the vertical up–down axis. The opposition between UP and DOWN, which is the topic of this paper, can be defined relative to gravitation, for example by observing the dropping of a stone, which under ideal conditions will go straight down to the ground. Less ideally, *up* and *down* are used also with reference to motion along a slanted plane (*The barrel rolled down the slope*). Many of the corpus examples analyzed below will be of this non-ideal type.

1.2 Aim and structure of the present study

The present study will focus on the use of *upp* and *ner* with a literal spatial meaning to express motion (non-literal uses will be discussed in a separate paper). Several types of motion situations will be distinguished. Motion typically refers to ***change-of-place*** (***displacement***) as in *Bert went from London to Berlin*, but motion verbs can also be used to describe ***motion within an area*** (***dynamic location***) as in *Bert strolled around in Berlin*. With respect to change-of-place, a major distinction can be made between ***subject motion***, when the entity that moves is realized as a subject as in *Bert was running*, and ***object motion*** as in *Bert put the vase on the table*, in which the motion of the object is profiled. When the motion verb is used in a sentence that refers to change-of-place, it is usually combined with an indication of the Path, which can be divided into three parts: Source – Transition – Goal (or Endpoint), as in *Ann went out of the forest* (Source), *up through the flowery meadow* (Transition) *and down to the lake* (Endpoint). Talmy (1985) also regarded stative ***Location*** as a type of motion situation: *Ann was down at the lake*.

The major research question is to answer what the exact functions of *upp* and *ner* are when they are used in descriptions of the motion situation and why in many cases they lack a direct counterpart as it turns out, not only in French but also in German and Finnish that are also satellite-framed. Related to this is the description of the crucial interplay between particles and prepositions in Swedish and English, and in German and Finnish with the addition of case. The role of the verb will also be studied since the different types of motion referred to by the verb have important consequences for the realization of Path even in a "satellite-framed" language such as Swedish.

The rest of the article is structured as follows. Section 2 will present data and method. Section 3 will discuss the use of *upp* and *ner* to describe subject motion followed by an account of Object motion in Section 4. The article ends with summary and discussion in Section 5.

2. Data and method

Data will be taken from the Multilingual Parallel Corpus (MPC), which consists of extracts from 22 Swedish novels and their translations into English, German, French and Finnish. There are around 600,000 words in the Swedish original texts. The source texts are indicated with a text code based on the author's name (see Appendix in Viberg, 2013b for a list of the Swedish originals and the text codes).

The choice of novels as texts will make it possible to study the basic spatial uses in a systematic way and to situate the result within a broader typological framework (the Talmian tradition). A broad range of non-spatial extended meanings are also represented in the corpus, but they will be accounted for in a separate paper. A study of the basic meanings is a necessary first step to construct a complete network of all the meanings.

The use of translations in contrastive studies is a bone of contention, but a thorough discussion would require a separate paper. Very briefly put, the advantage of using translation corpora is that the expression of the same meaning in the same context can be compared across languages. The problem is that you must take various translation effects into account. You can counteract that by including originals and translations from both languages if you compare two languages (see Viberg, 2016a and 2020, for my view) but for a single multilingual study such as the present one the inclusion of originals in all languages represents a vast undertaking even if it may in principle be followed up in later studies or by comparison with other research of the languages in question.

3. Subject motion

The use of *upp* och *ner* interacts with the types of Prepositional Phrases (PPs) and other spatial expressions that are used.

3.1 No PP

In the simplest case, there is no other indication of the Path than the particle. Looking at such structures is a good starting point to find the closest translational correspondences. When the particle is the only spatial satellite of the verb, the Goal must be inferred from the speech situation or from the discourse context. In example (1), the Goal (the attic) has already been mentioned in the preceding context. The context is shown only for the English version, which is representative of all the languages. (KE is an example of the text codes referred to in Section 2, INF3 = the third infinitive).

(1) The entrance hall was lined with pale-green
boarding and bright-blue wallpaper;
behind a door steep stairs led up to the attic.
She went ***up*** to look
Hon gick ***upp*** och tittade. KE — She went up and looked
Sie stieg ***hinauf*** und sah sich um. — She stepped DIST-up /---/
Elle ***monta*** vérifier. — She moved=up (to) check
Annie meni ***ylös*** katso-ma-an.
Annie went up look-INF3-Illative

The most frequent translations are listed in Table 1 together with some representative examples of less frequent translations.

Table 1. Major translations of Swedish *upp* when there is no PP.

Total Swedish *upp*: 82											
English			German			French			Finnish		
	n	%		N	%		n	%		n	%
up	53	65	(hin/her)auf-	39	48	(re)monter	12	15	ylös(päin)	21	26
			hoch-	12	15	se lever	11	13			
			(nach) oben	7	9	grimper	8	10	nousta 'rise'	14	17
Total[1]	53	65		58	71		31	38		35	43

The major English translation is *up*, but verticality is also signaled in the verbs *rise* and *climb* (7 + 6 tokens). In German, the most frequent translation is the separable particle *auf-* which is often preceded by one of the deictic markers *her-* (proximal) or *hin-* (distal, DIST). The particle has the same form as the preposition *auf* 'on'. (The hyphen on the particle is used to show that it is bound to the verb in certain contexts.) The second most frequent correspondence is *hoch-* which is basically used as an adjective *hoch* 'high' but can also be used as a separable particle. In Finnish, the most frequent translation is *ylös* (including 3 *ylöspäin* 'upwards'). *Ylös* is in the now obsolete lative case, which indicated direction to a goal (Kolehmainen, 2005: 136). It is related to other words such as *yli* 'over', *yllä* 'on, over'. The second-most frequent translation is a verb indicating vertical movement *nousta* 'rise', which in some cases is combined with *ylös*. A look at the most frequent translations shows that the use of the closest correspondence of *upp* varies: English *up* (65%), German (*hin/her*)*auf-* (48%) and Finnish *ylös* (26%). The correspondence is highest in English. As will be evident below, the correspondence is rather high in general for English when the particle has a literal spatial meaning, whereas the divergence is more pronounced between Swedish and the other languages.

As expected, there is no direct equivalent to *upp* in French, but vertical motion is primarily indicated in verbs. The major alternative is *monter* 'move=up', which, however, only accounts for 15% of the translations. There are other verbs, in particular *se lever* 'rise, lit. raise oneself', which also refers to vertical movement. *Grimper* 'climb up' is marked for manner but appears to indicate verticality as well ('move=up clambering'). None of these alternatives account for any large proportion of the translations, but they could be said to contribute to conceptual correspondence, since they indicate verticality. However, in many cases the translations into French lack any indication of verticality. Among verbs not marked for verticality, *arriver* 'arrive, come' (7 occurrences) and *venir* 'come' (4) are the most frequent. Like *upp*, these verbs indicate the reaching of the Goal but lack any indication of verticality. In a few cases, verticality is indicated in special phrases. For example, in one case, *arriver* is combined with *en haut* 'in high' and in one case with *au sommet* 'at/to the top'. Actually, there are other correspondents expressing UP besides the most frequent ones in all languages, but UP is not expressed in any way in many of the French and Finnish examples, so, to a great extent, correspondence is lacking even at the conceptual level.

Ner (with the alternative form *ned*) follows the same pattern as *upp*. In (2), *ner* is translated with a particle except in French, where the Path is incorporated in the verb *descendre* 'move=down'. In German, *unter-* 'down' is combined with the deictic prefix *her-*. In Finnish, *alas* is the most direct equivalent of *ner*. (CONDitional is a mood realized as a suffix: *-isi*.) Like *ylös*, *alas* is the obsolete lative form of a stem related to the noun *ala* 'area' and to spatial words such as *alla* 'under' and *alta* 'from beneath'.

[1] Percentages of the totals have been calculated separately and not by addition.

(2) Vad skulle hon säga när Mia kom ***ner***? (KE)
What should she say when Mia came ***down***?
Was sollte sie sagen, wenn Mia ***herunter***kam?
Qu'allait-elle dire quand Mia ***descendrait*** ?

Mitä	hän	sanoisi	kun	Mia	tulisi	***alas***?
What	(s)he	say-COND	when	Mia	come-COND	down

Table 2 shows the most frequent translations. The major French translations incorporate DOWN in a verb: *descendre* and *tomber*. (The forms *redescendre* and *retomber* with the repetitive prefix *re-* are included in the counts.) It must be remarked that when *tomber* is used, the Swedish original often contains *falla* 'fall'.

Table 2. The major translations of Swedish *ner/ned* when there is no PP.

Total Swedish *ner/ned*: 67											
English			German			French			Finnish		
	n	%		n	%		n	%		n	%
down	43	64	(he)runter-	18	27	descendre	22	33	alas	23	34
downstairs	2	3	hinunter-	13	19	tomber	11	16	alaspäin	3	4
			unter-	5	7						
Total	45	67	Total	36	54	Total	33	49	Total	26	39

English uses a structural equivalent most, followed by German and then Finnish, whereas the use of a Path-incorporating verb is the major strategy in French.

3.2 Combinations with the locative prepositions *på* and *i*

A specific characteristic of Swedish is that the two most frequent locative prepositions *i* 'in' and *på* 'on' are often used to indicate the endpoint of motion as in (3).

(3) När det blivit riktigt mörkt gick vi ***upp på*** vinden. (KÖ) *upp* 'up' + *på* 'on'/Prep
When it was properly dark we went ***up to*** the attic.
Als es richtig dunkel war, gingen wir ***auf*** den Speicher. *auf* 'on'/Prep + Acc
Ce soir-là, à la nuit tombée, nous ***montâmes au*** grenier. *monter* 'go=up' *à* 'at/to'

Kun ilta oli pimennyt kunnolla,	men-i-mme	***ylös***	vinti-***lle***
	go-PAST-1PL	up	attic-Allative ('onto')

In (3), *på vinden* (literally 'on the attic') indicates the endpoint of motion, but what does *upp* indicate? *Upp* indicates that the Endpoint is situated at a higher point than the point at which the motion started without indicating the resulting relation between the Figure and the Ground as the preposition *på* does. Thus *upp* describes the (final part of) the Path leading to the Endpoint of motion and will be referred to as the Trajectory for the purposes of this paper. As will be demonstrated, the reference to the Trajectory is often missing in the translations into German as in (3). In German, certain prepositions such as *auf* 'on' and *in* 'in' are combined with an NP in the accusative to indicate change-of-place and an NP in the dative to indicate Location and motion within an area. In Swedish, a directional particle such as *upp* is in principle obligatory to indicate change-of-place in combination with locative prepositions, whereas the use of a particle is optional in German. The use of a particle is optional also in Finnish, since Finnish has directional cases that indicate change-of-place: allative ('onto') and illative ('into') (see Huumo and Ojutkangas, 2006 for a more exact description).

Figure 1 shows the basic way of signaling motion UP and DOWN in Swedish in sentences with subject motion verbs.

Concepts	***Figure***	***Motion +Manner***	***Trajectory***	***Endpoint/ Location***	***Ground***
Phrase Structure	**NP**	**Verb**	**Particle**	**[Prep**	**NP]$_{PP}$**
	Linda Linda	gick went	upp up	på on	vinden the attic
	Lena Lena	sprang ran	ner down	i in	källaren the cellar
with directional preposition:					
	Lukas Lukas	åkte went	(upp) (up)	till to	tredje våningen the third floor (with the lift)

Figure 1. The basic way to signal motion UP/DOWN in Swedish with subject motion verbs.

At the conceptual level (top row), Path has been broken down into Trajectory and Endpoint and on the level of phrase structure (simplified) to Particle and Preposition. In the present context, Trajectory refers to the meanings UP/DOWN, but with some adjustments several particles with other meanings could be included in the scheme. Table 3 shows the major types of translations of *upp* in combination with the locative preposition *på* 'on'. Total refers to the expressions that express the concept UP. Zero translations of *upp* are indicated at the bottom of the table as explained below.

Table 3. The major translations of *upp* in combination with the locative preposition *på* 'on'.

Total Swedish *upp på*: 36											
English			German			French			Finnish		
	n	%		n	%		n	%		n	%
up on	7	19	(hin/her)auf-	8	22	monter	11	31	ylös(päin)	4	11
up to	8	22	hoch-	1	3	grimper	6	17			
up onto	5	14									
up (Other)	2	6									
Total	22	61		9	25		17	47		4	11
			Zero+PPacc	21	58				Zero+NPall	18	50
									Zero+NPill	5	14

Up appears in 61% of the English translations. One difference from Swedish is that *på* often corresponds to a directional preposition *to* or *onto*. *To* has a direct correspondent in Swedish *till* but that preposition would often sound unidiomatic in examples such as (3). The complex preposition *onto* lacks a Swedish correspondence. A possible explanation could be that *up* does not signal direction as clearly as *upp*, since *up* can have a locative meaning, which is covered by the specific locative form *uppe* in Swedish. In German, the closest correspondence (*hin/her)auf-* only occurs 8 times. Frequently, the endpoint is marked simply by *auf* as a preposition and change-of-place is indicated by the use of the accusative case, whereas there is a Zero-translation of *upp*. This alternative is symbolized: Zero+PPacc in Table 3. Even when there is a particle, there is usually also a PP but that is not indicated in this and the following tables.

In French, the most frequent translation is *monter* ‘move=up’, but it only occurs 11 times (31%). *Grimper* ‘climb’ indicates manner but also expresses motion in the vertical direction. In Finnish, the particle *ylös* ‘up’ occurs only four times in the translations. In most cases, the concept UP is not expressed (Zero), but direction is expressed simply by one of the directional cases. Allative is used most frequently since it roughly means ‘onto’, but illative, which roughly means ‘into’ is also used, since the correspondence is not perfect. (This difference is not relevant for the present paper.) These alternatives are symbolized Zero + NPall and Zero + NPill in Table 3. An NP marked with a directional case is usually present also when there is a particle, but this is not indicated in the tables.

Example (4) shows the most typical correspondences of *upp i*.

(4)	Lorelei klättrade ***upp i*** helikoptern	upp ‘up’/Particle + i ‘in’/Prep
	Lorelei climbed ***up into*** the helicopter	
	Lorelei kletterte ***in*** den Hubschrauber	in ‘in’/Prep + NPAccusative
	Lorelei ***monta dans*** l'hélicoptère	monter ‘go=up’ dans ‘in’
	Lorelei kipusi helikopteri***in***	Lorelei climbed helicopter-Illative

As can be observed in Table 4, *up* is often combined in English with the complex directional preposition *into*, which lacks a Swedish correspondent. The fact that *up* marks direction less clearly than *upp* may be at play even in this context. At a general level, the pattern is the same as for *upp på*. Verticality is in many cases Zero-translated in both German and Finnish and change-of-place is signaled by case, whereas French signals UP in the verb if the concept is signaled at all. Except for English, correspondence is low even at the conceptual level.

Table 4. The major translations of *upp* in combination with the locative preposition *i* ‘in’.

Total Swedish *upp i*: 34										
English			German			French			Finnish		
	n	%		n	%		n	%		n	%
up into	11	32	(hin/her)auf-	4	12	monter	14	41	ylös(päin)	0	0
up in	3	9	hoch-	1	3	grimper	3	9			
up (Other)	9	26									
Total	23	68		5	15		17	50		0	0
			Zero+PPacc	21	62				Zero+NPill	22	65
									Zero+NPall	4	12

Like *upp*, *ner/ned* often lacks a direct correspondence in the translations except in English, when the particle is combined with *på* or *i*, see Table 5 for *ner/ned på*.

Table 5. The major translations of *ner/ned* in combination with the locative preposition *på* ‘on’.

Total Swedish *ner/ned på*: 24										
English			German			French			Finnish		
	n	%		n	%		n	%		n	%
down on	9	38	(hin/her)unter-	5	21	descendre	2	8	alas	2	8
down to	2	8	nieder	2	8	tomber	3	13			
down onto	1	4									
down	2	8									
Total	14	58		7	29		5	21		2	8
Zero+onto	3	13	Zero+PPacc	11	46				Zero+NPall	11	46
									Zero+NPill	4	17

Motion down is special since all the MPC languages have path verbs that incorporate DOWN as part of their meaning. In (5), Swedish uses the verb *åka* 'move', which is completely neutral with respect to verticality. (Typically, this verb refers to motion as a Passenger in a vehicle, but in this example, *åka* implies that the motion was accidental.) The other languages except English use a verb that means 'fall' and thus signal motion down in the verb.

(5)	Ett papper hade åkt ***ner på*** golvet. (CL)	A paper had gone down on the floor
	A paper had fluttered ***to*** the floor.	
	Ein Blatt war ***auf*** den Boden gefallen.	A sheet was on the floor (acc) fallen
	Un papier était ***tombé*** par terre.	
	Yksi papereista oli pudonnut lattia***lle***.	One of (the) papers was fallen floor-Allative

The particle *ner* can also be combined with the preposition *i* as in (6).

(6)	Han kröp ***ner i*** hålet. (AL2)	down in/Prep
	He crept ***down into*** the hole.	
	Er kroch ***in*** das Loch.	in/Prep + Acc
	Il est ***descendu dans*** le trou.	moved=down in/Prep
	Hän ryömi kaivanto***on***.	(s)he crept ditch-Illative

The translations of the combination of *ner* + *i* 'in' follows the same pattern as *ner* + *på* (see Table 6), but it should be noted that *ner* + *i* is much more frequent.

Table 6. The major translations of *ner/ned* in combination with the locative preposition *i* 'in'.

Total Swedish *ner/ned i*: 82											
English			German			French			Finnish		
	n	%		n	%		n	%		n	%
down into	22	27	hinunter-	7	9	descendre	15	18	alas	5	6
down to	9	11	nach unten	5	6	tomber	16	20			
down in	5	6									
down	9	11									
Total	45	55		12	15		31	38		5	6
Zero+into	20	24	Zero+PPacc	55	67				Zero+NPillat	62	76
Zero+to	3	4							Zero+NPallat	7	9

A complicating factor in many of the examples is that there are occurrences of the relatively frequent path verbs *sjunka* 'sink' and *falla* 'fall' in the Swedish original texts. In addition, there are single occurrences of several more specific verbs that incorporate the concept DOWN in the verb (*ramla*, *trilla*, *dimpa*, *dråsa*, *droppa*, *störta*, *rasa*). Together, the path verbs indicating motion DOWN account for 21% (59/282) of the examples of V + *ner*.[2] When the path verb is combined with a locative preposition, the particle is not obligatory to signal change-of-place but *ner/ned* is often used, even if this may appear to be redundant. It would be too complicated

[2] Verbs incorporating DOWN appear to be common across languages. Viberg (2006: 113) suggested that the incorporation of Path in the verb root is variable and can be described as a markedness hierarchy, where verbs incorporating DOWN represent the first, unmarked step.

to account for all combinations that occur. In Table 5 and 6, the use of verbs of falling is not shown, except for French since the verb is the only way of indicating motion down in French.

På and *i* account for a large proportion of the combinations of Particle + Preposition. Except for the two directional prepositions *till* 'to' and *mot* 'towards' that will be discussed in the next section, *upp* is combined with 16 prepositions other than *i* and *på*, but they reach at most a frequency of 5 occurrences except *ur* 'out of' (13 occurrences). *Ner/ned* is combined with 12 other prepositions but only two appear more than five times: *från* 'from' (10) and *över* 'over' (8).

3.3 Combinations with directional prepositions

Swedish also has prepositions that express direction by themselves. When they are used, a particle is not needed to signal change-of-place. Such examples will therefore be treated rather briefly.

3.3.1 *The reaching of a Goal indicated by till 'to'*

The most basic directional preposition *till* 'to' refers to the reaching of a Goal, see (7).

(7) Sedan lämnade han strandboden och gick ***upp till*** huset. (MF)
Then he left the boathouse and walked ***up to*** the house.
Dann verließ er das Strandhaus und ging ***hinauf zum*** Haus.
Puis, laissant le cabanon, il ***monta jusqu'à*** la Maison de Verre.
Sitten hän poistui rantavajalta ja meni ***talolle***. [went house-Allative]

When *till* is used, no particle is obligatory in Swedish to mark the reaching of the goal, but *upp* and *ner* can be used to indicate motion in the vertical dimension as in (7). There are 43 occurrences of *upp + till* and 55 of *ner/ned till*.

3.3.2 *Motion towards a Goal indicated by mot 'towards'*

The preposition *mot* 'towards' expresses motion towards a Goal without indicating the reaching of the Goal, see (8).

(8) Och hon gick ***upp mot*** det gröna huset. (POE1)
And she walked ***up towards*** the green house.
Und sie ging ***zu*** dem grünen Haus ***hinauf***.
Et elle s'avançait ***vers*** la maison verte.

Ja	hän	lähti	***nouse-ma-an***	kohti	vihreä-tä	talo-a
and	(s)he	left	rise-3INF-Illative	towards	green-Partitive	house-Partitive

When *upp* is combined with *mot*, the resulting clause refers to an activity and is atelic. This can be tested by adding a durational adverb: *Hon gick upp mot huset **i** 10 minuter / She walked up towards the house **for** 10 minutes*. In addition, *upp* and *up* are not stressed as a particle and function as spatial adverbs rather than as verbal particles. The same applies to *ner* in combination with *mot*. The total number of occurrences of *upp + mot* is 27 and of *ner/ned mot* 18.

3.4 Summing up subject motion

At a general level, the degree of structural correspondence between Swedish and the other languages is summed up in Table 7 by looking at the extent to which the closest structural equivalents are used across all examples of subject motion. (For French, there is no direct structural equivalent, but the most frequent translation is shown instead.)

Table 7. The major structural equivalent of *upp* 'up' and *ner/ned* 'down' across all types of subject motion.

	Swedish	English	German	French	Finnish
	upp	up	*auf-	monter	ylös
N	275	185	96	62	32
%	100	67	35	23	12
	ner/ned	down	*unter-	descendre	alas
N	282	177	100	75	42
%	100	63	35	27	15

There is a sharp contrast between the three S-languages in spite of the fact that they all have a structural equivalent. English uses *up* and *down* as translations almost twice as often as German uses **auf-* and **unter-*. (The star indicates the various deictic markers.) The Finnish particles are used even less. As noted, their grammatical status is unclear, but there are no other conceptually corresponding elements in most examples. The difference is not confined to the choice of structural elements to express verticality but also represents an important conceptual difference, since in most cases verticality is not referred to in any other way. The proportion of Zero translations is particularly clear when change-of-place is indicated only by case in the translations of Particle + *på*/*i* into German and Finnish, see Table 8.

Table 8. Zero translation of *upp* and *ner* in constructions of the type Particle + *på*/*i*.

	Swedish	German	Finnish
	upp på/i; ner på/i	Zero + PPacc	Zero + NPallative/illative
N	176	108	133
%	100	61	76

The frequent use of a few subject-motion verbs is a conspicuous feature (see Table 9) that needs to be commented on. Even though there are as many as 76 verb types combined with *upp* and 77 combined with *ner*, the six most frequent verbs account for around 50% of the verb tokens. This reflects a general tendency for a few verbs to dominate within their semantic fields in terms of frequency of occurrence. In the Swedish SUC corpus comprising around one million words from a variety of written registers, the two most frequent motion verbs, the nuclear verbs *komma* and *gå*, cover together more than 25% of the total number of motion verb tokens in the corpus (including all uses, not just the literal ones), and the ten most frequent motion verbs account for close to 50% (Viberg, 2006: 114).

Table 9. The most frequent subject-motion verbs used in combination with *upp* and *ner* in the MPC.

Upp			*Ner/ned*		
Total:					
Tokens		275			282
Types		76			77
gå	'go, walk'	57	gå	'go, walk'	50
komma	'come'	41	komma	'come'	24
klättra	'climb'	21	falla	'fall'	19
fara	'go/travel'	13	sjunka	'sink'	17
flyga	'fly'	9	krypa	'creep'	13
åka	'go, ride in a vehicle'	9	åka	'go, ride in a vehicle'	12
		150			135

The results can be compared to the study by Olofsson (2018), who looked at 17 constructions consisting of a verb followed by a particle and a PP, for example [V-*runt*-på] and [V-*upp*-*till*]. The type of motion was restricted to what I call subject motion. The primary focus of Olofsson's study was the productivity of the constructions measured as the type and token frequencies of the verbs. In this study, the combined frequency of the ten most frequent verbs amounted to 71%, and the two most frequent verbs were *gå* and *komma* followed by *åka* 'go, ride in a vehicle'. The study was based on large samples from corpora, in total 22 978 verb tokens (Olofsson, 2018, Table 5, p. 189).

3.5 Non-vertical interpretations of *upp* and *ner*

As discussed in Strzelecka (2003: 223–227), *upp* and *ner* do not always refer to motion along the vertical dimension when they refer to change-of-place. No attempt will be made to quantify all such cases systematically, since the justification of their interpretation often requires detailed discussion. One specific use that is relatively frequent in the MPC is ***geographic up/down***. In Swedish, *upp* can refer to a place that is further north than the starting point (see 9). In a similar way, *ner* can refer to motion towards the south.

(9) Det var ju därför Rebecka åkte ***upp till*** Kiruna,
That's why Rebecka went ***up to*** Kiruna,
Ja, und deshalb ist Rebecka ***nach*** Kiruna gefahren,
C'est pour ça que Rebecka est partie ***à*** Kiruna,
Sen takiahan Rebecka oli lähtenyt Kiiruna***an***, (Kiruna-Illative]

Table 10 accounts for the translation of geographic *upp* in descriptions of subject motion. As can be observed, only English uses the direct correspondent *up* as a translation more than in a few sporadic examples. In Finnish, the north is referred to in some translations (*pohjoiseen* 'north-Illative').[3]

Table 10. Translations of Swedish geographic *upp.*

Total Swedish geographic *upp*: 20							
English		German		French		Finnish	
up	14	herauf-	2	(re)monter	2	ylös	1

[3] Geographic north is described in a Swedish grammar (Viberg *et al*., 1984 §15.8), which has been translated into all MPC languages. The translators in various ways corroborate the claim that geographic *upp* lacks a correspondent in French and Finnish. An example is given of the locative form *uppe*: *Sommaren tillbringar de uppe i fjällen*. *They spend the summer up in the mountains. Il passent l'été (au nord) dans les montagnes*.

Upp and *ner* can also refer to ***the center of attention***. A representative example is found in (10). (The English translation is quoted first to show the surrounding context.)

(10) The site was crowded with people. *Cars **drove up** to the office* and the people who got out were handsomely dressed, but Annie could see no faces, only eyes.
*Bilar **körde upp** framför expeditionen.* KE [Cars drove up in-front-of the office]
Autos ***fuhren*** vor der Anmeldung ***vor***.
Des voitures ***se garèrent*** devant la réception.
[Cars were parked in-front-of the office[4]]
Autoja ***ajoi*** toimiston eteen. [Cars drove office-GEN front-Illative]

It appears that "the office" referred to in the English version is not situated in a high location that would motivate the use of *up*. (As is often the case, it is difficult to be sure about the intended interpretation, but in any case, it is possible to give the example a natural interpretation under the assumption that no vertical displacement is involved.) The scene is described through the eyes of Annie. First a general view: *The site was crowded with people*. Then her attention is drawn to cars driving up and after that the focus narrows down to the people getting out of the cars and finally to an attempt to focus on their faces. German uses a particle (the final *vor*) that refers to motion forward, which is often used to indicate center of attention.[5] (This will be studied in a separate article on the Swedish particle *fram* 'forward'.) Several other uses where *upp* refers to horizontal motion are discussed in Strzelecka (2003: 223–226) and in Ekberg (1997), but each of them has a low frequency in the corpus, which makes it difficult to make a contrastive comparison. At sea, for example, *upp* can refer to motion toward the wharf as in *han girade upp mot bryggan* 'he steered [up] towards the wharf' (Center) or to motion against the wind as in *vända upp i vind* 'turn upwind'. As in English, you go up and down a stream. Both languages also have similar compound adverbs: *upstream-downstream*; *uppströms-nerströms*).

4. Object motion

The use of *upp* and *ner* to refer to object motion follows a similar pattern as for subject motion except that the interaction with different types of verbs is more complex. For that reason, some of the conclusions in this section will be tentative, but it is important to include object motion into the analysis since various types of caused motion is an understudied area, as pointed out by Matsumoto and Kawachi (2020). The major focus in Section 4 will be on phenomena that distinguish object motion from subject motion.

4.1 The general picture

The major types of translations are the same for object motion as for subject motion, but as will be discussed below, there are several types of object motion verbs in Swedish that do not require a particle to signal direction. In (11), which shows the general pattern, *upp* is translated by *up* in English, whereas the trajectory is incorporated into the verb in French. *Monter* 'move=up' can be used both as a transitive and as an intransitive verb. In German, direction is

[4] The verb *garer* '(to) park' is in the reflexive form (*se garer*), which makes the verb intransitive.
[5] French has a preposition (*devant*) and Finnish a postposition (*eteen*) which correspond to the Swedish preposition *framför* 'in front of' and not to *upp*.

signaled by the accusative case and in Finnish by the allative case (‘onto, to’), and the concept UP is not expressed in any way.

(11) Henry hade ***baxat upp*** TV:n ***på*** vinden (KE)
[Henry had lugged up the TV on the attic]
Henry had ***lugged*** the television ***up*** there
Henry hatte den Fernseher ***auf*** den Boden ***geschleppt***
Henry avait ***monté*** la télévision ***au*** grenier
Henry oli ***raahannut*** television vinti***lle***
[Henry had lugged TV attic-Allative (‘to’)]

The schema used for subject motion in Section 3.2 can be modified to describe object motion simply by inserting an object slot directly after the particle, see Figure 2.

Agent	Motion +Manner	Trajectory	Figure	Endpoint	Ground
NP	Verb	Particle	NP	Prep-loc	NP
Henry Henry	bar carried	upp up	teven the telly	på on	vinden the attic
Ann Ann	ställde put (‘stood’)	(ner) down	väskan the suitcase	på on	golvet the floor

Figure 2. The signaling of Path in object motion in Swedish.

Table 11 accounts for the most frequent correspondences of all categories of *upp* in combination with an object motion verb (both *upp* + no PP and *upp* + all types of PPs). Only the elements that refer to verticality are included.

Table 11. The major translations of *upp* in descriptions of Object motion.

Total Swedish *upp*: 195											
English			German			French			Finnish		
	n	%		n	%		n	%		n	%
up	117	60	(hin)auf-	31	16	(re)monter	20	10	ylös	16	8
upstairs	1		hoch-	38	19	soulever	21	11	ylöspäin	1	1
upright	1					hisser	9	5	pystyyn	6	3
upward	1					dresser	3	2			
Total:	120	62		69	35		53	27		23	12

Starting with English, it can be observed that *up* is a rather frequent translation of *upp*, whereas the closest German correspondent (*hin*)*auf-* accounts for only a small proportion, and even together with *hoch-* ‘high’ used as particle, vertical movement is expressed only in 35% the examples. In Finnish, *ylös* accounts for a still smaller proportion. There are a few other markers such as *pystyyn* ‘upright’, but in most cases verticality is not expressed. As with Subject motion, Zero translations of UP are frequent in German and Finnish and direction is signaled via the case system. In French, several verbs incorporating UP are used, but UP is rather frequently not expressed. The most frequent subject-motion verb *monter* ‘move up’ is used also as an object-motion verb, but other verbs referring to vertical movement are also used such as *soulever* ‘lift’, *lever* ‘lift, raise’, and *hisser* ‘hoist’ and *dresser* ‘raise’. Actually, verbs of the latter type are used in all the MPC languages as can be observed in (12).

(12) Sedan ***lyfter*** han ***upp*** hunden och kramar om den. (PCJ2) [Then lifts he up the dog]
Then he ***lifts up*** the dog and embraces it.
Dann ***hebt*** er den Hund ***hoch*** [Then lifts he the dog high]
Puis il ***soulève*** le chien [Then he lifts the dog]
Sitten hän ***nostaa*** koiran ***syliinsä*** [Then he lifts dog bosom-Illative-3Poss]

One thing that complicates the picture is the fact that some of the verbs that are used in the Swedish originals refer to motion in the vertical dimension and could be said to incorporate UP in their most unmarked use. The most frequent of these verbs in the present material is *lyfta* 'lift' (25 occurrences), followed by *hissa* 'hoist' (5), but both *lyfta ner* 'lift down' and *hissa ner* 'hoist down' are possible, and examples of these combinations appear in the MPC. It is open to discussion, whether a verb such as French *soulever* 'lift' should be counted as a correspondence of *upp,* even in the cases when the Swedish original contains *lyfta*. On the other hand, it is not quite satisfactory to count this as a case of Zero translation of *upp,* since the concept UP is present also in the French version. The situation in Finnish is also complex. In Finnish, the most frequent translation of *lyfta* 'lift' is *nostaa,* which is used 35 times as a translation, but only once in combination with *ylös* 'up' (and two with *pystyyn*). *Nostaa* (and its derived forms causative *nostattaa* and frequentative *nostella*) is used also as a translation of some non-directional Swedish verbs combined with *upp*. These and similar complications must be left for studies based on all occurrences of specific object-motion verbs in a corpus rather than on the particles.

The use of *ner/ned* 'down' in combination with a motion verb follows the general pattern and will not be commented on in detail. See Table 12.

Table 12. The major translations of *ner/ned* in descriptions of Object motion.

Total Swedish *ner/ned*: 183											
English			German			French			Finnish		
	n	%		n	%		n	%		n	%
down	91	50	her-/r-/hinunter	36	20	descendre	9	5	alas	26	14
			(hin-)ab	17	9	baisser	7	4	alaspäin	4	2
			nieder-	3	2	plonger	11	6			
			nach unten	2	1				maahan	6	3
Total	91	50		58	32		27	15		36	20
			Zero+PPacc	73	40				Zero+NPill	68	37
									Zero+NPall	24	13

In French, *descendre* 'move=down' can be used also as an object-motion verb, but *baisser* 'lower' and *plonger* 'dive, plunge' are also used approximately as frequently in the present material. The closest Finnish correspondents are *alas* 'down' and *alaspäin* 'downwards' do not reach a very high frequency. *Maahan* 'to the ground' [ground-Illative] has been included in the table as an example of a general noun inflected for case and used as a correspondence of *ner*. Zero translations of *ner/ned* and the use of case to signal change-of-place are frequent in German and Finnish also with this particle.

4.2 The decisive role of object motion verbs

Rather many different verbs are used to describe object motion but as can be observed in Table 13 the most frequent verbs account for a large proportion of all verbs.

Table 13. The most frequent object-motion verbs used in combination with *upp* and *ner* in the MPC.

upp			*ner/ned*		
Total tokens		195			183
Total types		59			78
lyfta	'lift'	25	dra	'pull'	20
dra	'pull'	20	stoppa	'put=stuff'	14
ta	'take'	10	lägga	'lay'	13
plocka	'pick'	8	sticka	'put=stick'	11
lägga	'lay'	9	släppa	'let go'	10
sätta	'put=attach'	9	köra	'drive'	8
			ställa	'put=stand'	7
Total		81			83

A special group that is characteristic only of *upp* is the verbs of taking (*ta* 'take', *plocka* 'pick' and *fiska* 'fish, acquire'). Most of the other verbs belong to basic groups of object motion verbs in Swedish. Verbs of pulling and pushing are represented by *dra* 'pull' in Table 13. These verbs in general require a particle to express change-of-place in combination with locative prepositions as in (13), which describes the pulling of a canoe out of the water onto the grass on the shore.

(13) När han skulle ***dra upp*** den ***i*** gräset (KE)
As he was about to pull the canoe ***up on*** the grass,
Als er das Kanu ***ins*** Gras ziehen wollte,
Au moment où il allait le ***tirer sur*** l'herbe,
Kun hän ***veti*** kanoottia ruohikko***on***, [when (s)he pulled canoe grass-Illative]

A sentence without the particle *upp* such as in *Han drog kanoten i gräset* 'He pulled the canoe in the grass' rather describes an atelic situation where the canoe already is in the grass at the start of the motion. Another basic group of object-motion verbs that require a particle are verbs of carrying, such as *bära* in sentences like *Ann bar upp resväskan på vinden*, literally 'Ann carried up the suitcase on the attic'. Without the particle, this sentence would rather mean that Ann carried the suitcase around on the attic or sound odd. As discussed in Viberg (2015a), verbs of carrying and verbs of pulling and pushing describe co-motion of agent and object, which makes it possible to use such verbs to refer to motion within an area.

Another frequent type of object-motion verbs are the verbs of putting such as (in Table 13) *lägga* 'lay', *sätta* 'put-attach', *ställa* 'put-stand', *stoppa* 'put-stuff' and *sticka* 'put-stick'. Such verbs frequently identify the endpoint of motion with *på* and *i* without using a particle, see (14).

(14) Han hade ställt en panna ***på*** spisen. (POE)
He had put a pan ***on*** the range.
Er hatte einen Topf ***auf*** den Herd gestellt. auf/Prep + NPacc
Il avait posé une poêle ***sur*** la cuisinière.
Hän oli nostanut hella***lle*** pannun. [(S)he was lifted stove-Allative pan-ACC]

Actually, verbs of putting are frequently combined with locative prepositions in Swedish without using a particle (Viberg, 2015b: 230), and in that case Swedish looks like French, but it would be wrong to say that Swedish behaves like a V-language, since only change-of-place is indicated in the verb and not a specific Path, and unlike a typical V-language such as French,

the Trajectory can easily be specified by adding a particle. Change-of-place of the object is rather a defining feature of a verb of putting that is present in all languages that have such verbs. German and Finnish do not treat verbs of putting separately but signal direction by case in the same way as with other verbs.

4.3 A survey of the use of particles to represent the endpoint of motion in Swedish

The major strategies used to express the reaching of the endpoint of motion are summed up in Table 14. (There are also other types of markers that have not been discussed in this paper, most importantly deictic adverbs: *hit* 'hither' and *dit* 'thither' and question words *vart* 'where to').

Table 14. The expression of endpoint of motion in Swedish.

To signal directional motion (Change-of-place) Swedish uses:

(i) Canonical Satellite-framed patterns:

- A directional particle alone: *Peter gick upp.* 'Peter went up'
- Directional particle+ Locative preposition: *Peter gick upp på vinden.* 'Peter went up to the attic'
- (Particle) + Directional Prep: *Peter gick (upp) till huset på kullen.* 'Peter went (up) to the house on the hill'

(ii) Non-canonical patterns: (Particle) + Locative preposition

- Verbs of falling: *Peter trillade (ner) i vattnet.* 'Peter fell (down) in(to) the water'
- Verbs of throwing: *Peter kastade boken i papperskorgen*. 'Peter threw the book in(to) the wastepaper basket'
- Verbs of putting: *Peter la boken på bordet.* 'Peter put the book on the table'
- Verbs of pouring: *Peter hällde vatten i flaskan*. 'Peter poured water in(to) the bottle'

The canonical satellite-framed patterns require that all elements that indicate the Path appear outside the verb as satellites of some type. In Swedish, directional particles such as *upp* and *ner* play a crucial role to indicate change-of-place unless there is a directional preposition such as *till* 'to'. Verb-framed patterns are non-canonical in Swedish, but there are a number of path verbs. The most important group is the verbs of falling that clearly incorporate DOWN in their meaning as in a V-framed language. It should be noted that they rather often are combined with the particle *ner/ned* in spite of the fact that verticality is already indicated by the verb. Obviously, regularization also plays a role. Since particles are required in many other cases, there is a tendency that the combination Particle + Preposition is used whenever it can be used. Among the object motion verbs there are also verbs that refer to vertical motion, but as discussed above, many of these verbs require separate analysis since UP appears to represent only a default in verbs such as *lyfta* 'lift' and *hissa* 'hoist'. There are also a number of object motion verb such as the verbs of putting that do not indicate any particular Path but nevertheless indicate change-of-place of the object when combined with a locative PP without requiring a directional particle. These verbs share the characteristic that they do not refer to co-motion of

the Agent. German and Finnish do not treat such verbs differently but indicate change-of-place explicitly by case in the ordinary way. Verbs that presuppose co-motion of the Agent such as the verbs of carrying and the verbs of pulling and pushing require a directional particle in Swedish to avoid that the sentence is interpreted as motion within an area.

4.4 Pouring: a case of reconceptualization across languages

In Swedish, pouring fits into the general picture of object motion but was not included in the count presented in Table 11. Pouring will be presented separately in this section, since the translations into the other languages to a great extent are based on different types of metonymy or a reconceptualization of the motion event as another type of event. Pouring basically refers to the caused motion of liquid from one container_1 held in the hand into another container_2 by tilting container_1. This basic conceptualization is shared by all the languages to be discussed below. However, a complex situation is often referred to by simply mentioning some salient subevent and leaving to the interpreter to fill out the rest. This is a kind of metonymy (subevent for event, pars pro toto). The Swedish sentence *Åke hade hällt upp whisky* [Åke had poured up whiskey] in (15) refers to the rising of the liquid in container_2 without explicitly mentioning any container.

(15) Åke hade ***hällt upp*** whisky och satt fram knäckebröd och öl. (KE)
Åke had ***poured out*** whisky and put out crispbread and beer.
Åke hatte Whisky ***eingeschenkt***
Åke avait ***versé*** du whisky
Åke oli ***kaatanut*** viskiä ***laseihin***

Languages differ with respect to what elements they choose to express explicitly. In English, *pour out* refers to the result of tilting container_1 and the rest must be inferred. In both German (*einschenken* ‘in-serve’) and Finnish (*kaataa lasei-hin* ‘pour glasses-Illative), reference is made to the motion of the liquid into container_2, which is explicitly mentioned in Finnish but understood in German. In French, only the liquid is explicitly mentioned (*verser du whisky* ‘pour whiskey’). See Table 15. In Swedish, the verb *slå* ‘hit’ is frequently used as a (near) synonym to *hälla* ‘pour’ and appears in the same constructions (Viberg, 2016b: 207–208).

Table 15. The major translations of *hälla upp*/ *slå upp* (a liquid) in the MPC.

Total Swedish *hälla upp*/*slå upp*: 19							
English		German		French		Finnish	
pour out	5	einschenken	9	verser	10	kaataa + Zero	10
pour + Zero	13	eingiessen	2	servir	7	kaataa + NPillative	6
fill	1	auffüllen	1			kaataa + NPallative	2

In English, *pour* is used in combination with *out* in 5 examples, but in most cases it is used without any spatial particle. In German, *ein-* ‘in’ appears in the majority of the translations, but *auf-* ‘up’ appears in one example (*auffüllen* ‘fill up’). French uses two verbs: *verser* ‘pour’ and *servir* ‘serve’ but no spatial markers. The latter represents a reconceptualization of the situation as serving something to someone (cf. German *einschenken*). Finnish uses the verb *kaataa* ‘pour’. In 8 cases container_2 is indicated as a Goal with a directional case as the illative (*-hin*) in (15). The Swedish use of *upp* ‘up’ to indicate the rising of the liquid in container_2 appears to be very language specific.

5. Summary and discussion

The study of the typological profile of a language differs from an ordinary contrastive study in several respects. It is multilingual, and it has as an aim to identify language-specific features and features that are characteristic of the language's general type in some sense. As stated in the beginning of this paper, the profile is based on work in general typology and contrastive studies. The first step is to identify the place of the structures being studied in a general typology. Spatial language has already been studied extensively from a typological point of view. In a broad survey of grammars of space in the languages of the world, Levinson and Wilkins (2006: 527) conclude that most languages that can be assigned to a certain type are Verb-framed and that the Germanic Satellite-framed pattern may be very restricted typologically. Thus, saying that Swedish is a satellite-framed language means that it differs in this respect from most of the world's languages.

The present study zooms in on four S-languages and points to important intra-typological differences with respect to the expression of spatial relations in the description of motion events.

- Swedish exploits directional particles such as *upp* and *ner* to signal change-of-place in combination with locative prepositions such as *på* and *i*. There is a clear difference between Swedish and each of the other S-languages. English translations tend to contain *to* (*into*, *onto*) instead of a locative preposition. German frequently uses only a prepositional phrase and uses case to signal the distinction between location and change-of-place. Finnish frequently uses only an NP marked with a directional case. French, the only V-language in this study, as expected uses path verbs that incorporate UP or DOWN, but in many cases these two concepts are not expressed explicitly.

- The use of directional particles in Swedish to signal change-of-place leads to conceptual differences. The trajectories UP and DOWN are expressed more frequently to describe motion events in Swedish than in German, Finnish and French. For English, the situation is less clear, since the correspondent particles are used frequently, and it should not be expected that the closest correspondence is always used, even when languages are similar on a certain point.

Verbs play a prominent role to signal spatial relations (not just manner) even in an S-language.

- To begin, path verbs, in particular verbs indicating uncontrolled motion DOWN are relatively frequent in all the languages studied. This may be a universal tendency. There are also some verbs indicating motion UP in all the S-languages, but variation between such languages requires further study.

- In Swedish, several basic subfields of object motion verbs such as verbs of putting do not require spatial complements that signal change-of-place but often use complements that simply indicate location. German and Finnish in general use complements marked for direction also in this case.

- There are several more subfields of motion verbs that require separate study to give a full picture of the description of motion events. See Matsumoto and Kawachi (2020) for other such studies.

- A general conclusion that can be drawn is that the indication of the endpoint of motion or final location (often signaled with a PP or, in Finnish, a case-marked NP) is more important than the signaling of the general orientation of the motion (the trajectory).

The original plan for this article was to cover all major uses of *upp* and *ner*, but for reasons of space, the account of non-spatial extensions will be published as a separate article that will also include a discussion of compound forms, in particular compound verbs, which represent another language-specific characteristic of Swedish (for a general overview, see Viberg, 2017). Another limitation of this study has to do with the multilingual approach which focuses on Swedish. The other languages have been discussed in less detail, but I feel confident that the broad differences that have been identified will hold and hope that the characterization of Swedish in this paper can be used to give a perspective on the other languages if they are analyzed in depth in other studies.

I will conclude by discussing some of the applications of the description of the typological profile of a language such as bilingualism and translation. Information about various areas of the typological profile of Swedish guided my own work on a Swedish grammar for second-language learners. My earlier contrastive-typological studies have been carried out in parallel with work on the acquisition of Swedish as a second language. A general conclusion from that work was that all learners tended to have problems acquiring language-specific features. In addition, there were problems characteristic of specific source languages. This calls for the combination of a typological and a contrastive approach. (cf. Filipović, 2017 on "applied language typology"). The situation in Sweden, where recent immigrants with many diverse first languages are often taught together in the same group, has parallels in many parts of the world. It will often not be possible to gain access to detailed contrastive descriptions of each language involved.

References

Aurnague, M. and Stosic, D. 2019. Recent Advances in the Study of Motion in French: A Survey. In M. Aurnague and D. Stosic (eds), 2–28.

Aurnague, M. and D. Stosic (eds). 2019. *The Semantics of Dynamic Space in French: Descriptive, Experimental and Formal Studies on Motion Expression*. Human Cognitive Processing 66. Amsterdam: Benjamins.

Beavers, J. Levin, B. and Tham, S.W. 2009. The Typology of Motion Expressions Revisited. *Journal of Linguistics* 46(2): 331–377.

Blomberg, J. 2014. Motion in Language and Experience: Actual and Non-actual Motion in Swedish, French and Thai. *Travaux de l'Institut de Linguistique de Lund, 53*. PhD thesis. The Faculties of Humanities and Theology.

Boers, F. 1996. *Spatial Prepositions and Metaphor. A Cognitive Semantic Journey along the UP-DOWN and the FRONT-BACK Dimensions*. Tübingen: Gunter Narr.

Cappelle, B. 2005. Particle Patterns in English. A Comprehensive Coverage. PhD thesis. Katholieke Universiteit Leuven.

Dehé, N. 2015. Particle Verbs in Germanic. In *Word-formation: An international handbook of the languages of Europe Vol. 1,* P.O. Müller, I. Ohnheiser, S. Olsen and F. Rainer (eds), 611–626. Berlin/Boston: De Gruyter Mouton.

Ekberg, L. 1997. The Mental Manipulation of the Vertical Axis: How to Go from "up" to "out", or from "above" to "behind". In *Lexical and syntactical constructions and the construction of meaning*, M. Verspoor, K.D. Lee and E. Sweetser (eds), 69–87. Amsterdam and Philadelphia: Benjamins.

Fagard, B. and Kopecka, A. 2021. Source/Goal (A)symmetry. A Comparative Study of German and Polish. *Studies in Language* 45(1): 130–171.

Fagard, B., Zlatev, J., Kopecka, A., Cerruti, M. and Blomberg, J. 2013. The Expression of Motion Events: A Quantitative Study of Six Typologically Varied Languages. *Berkeley Linguistics Society 39*: 364–379.

Filipović, L. 2017. Applied Language Typology: Applying Typological Insights in Professional Practice. *Languages in Contrast* 17(2): 255–278.

Gries, S. Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. [Open Linguistics Series] London: Continuum.

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

Huumo, T. and Ojutkangas, K. 2006. An Introduction to Finnish Spatial Relations, Local Cases and Adpositions. In *Grammar from the Human Perspective: Case, Space and Person in Finnish*, M-L. Helasvuo and L. Campbell (eds), 11–29. Amsterdam: Benjamins.

Ibarretxe-Antuñano, I. (ed.). 2017. *Motion and Space across Languages*. Amsterdam: Benjamins.

Kolehmainen, L. 2005. *Präfix- und Partikelverben im deutsch-finnischen Kontrast*. PhD thesis. Helsinki University.

König, E. 2012. Contrastive Linguistics and Language Comparison. *Languages in Contrast* 12(1): 3–26.

Lebas, F. and Cadiot, P. 2003. *Monter* et la constitution extrinsèque du référent." *Langages* 150: 9–30.

Levinson, S.C. and Wilkins, D. 2006. Patterns in the Data: Towards a Semantic Typology of Spatial Description. In S.C. Levinson and D. Wilkins (eds), 512–552.

Levinson, S.C. and Wilkins, D. (eds). 2006. *Grammars of Space. Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

Lindner, S.J. 1981. *A Lexico-semantic Analysis of English Verb Particle Constructions with* out *and* up. Bloomington, IN.: Indiana univ. linguistics club.

Luo, H. 2019. *Particle Verbs in English. A Cognitive Linguistic Perspective*. Springer Verlag.

Matsumoto, Y. and Kawachi, K. 2020. Introduction. Motion Event Descriptions in Broader Perspective. In Y. Matsumoto and K. Kawachi (eds), 1–22.

Matsumoto, Y and Kawachi, K. (eds). 2020. *Broader Perspectives on Motion Event Descriptions* Human Cognitive Processing 69. Amsterdam: Benjamins.

Olofsson, J. 2018. *Förflyttning på svenska*. (Motion in Swedish – on productivity from a construction grammar perspective.) Göteborgsstudier i nordisk språkvetenskap 32. PhD thesis. University of Gothenburg.

Slobin, D.I. 1996. Two Ways to Travel. Verbs of Motion in English and Spanish. In *Grammatical Constructions: Their Form and Meaning*, M. Shibatani and S.A. Thompson (eds), 195–220. Oxford: Clarendon Press.

Slobin, D.I. 2004. The Many Ways to Search for a Frog: Linguistic Typology and the Expression of Motion Events. In *Relating Events in Narrative: Typological and Contextual Perspectives*, S. Strömqvist and L. Verhoeven (eds). 219–257. Mahwah, NJ: Lawrence Erlbaum Associates.

Slobin, D.I. 2017. Typologies and Language Use. In I. Ibarretxe-Antuñano (ed.), 419–446.

Strzelecka, E. 2003. *Svenska partikelverb med in, ut, upp och ner. En semantisk studie ur kognitivt perspektiv*. [Swedish phrasal verbs with *in*, *ut*, *upp* and *ner*. A semantic study from a cognitive perspective.] Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet 62. PhD thesis. Uppsala University.

Talmy, L. 1985. Lexicalization Patterns: Semantic Structure in Lexical Forms. In *Language Typology and Syntactic Description*. Vol. 3, T. Shopen (ed.), 57–149. Cambridge: Cambridge University Press.

Talmy, L. 2000. *Toward a Cognitive Semantics*. Vol. II. Cambridge, MA: MIT Press.

Teleman, U., Hellberg, S., and Andersson, E. (eds). 1999. *Svenska Akademiens grammatik. Del 2. Ord.* Stockholm: Norstedts.

Van derAuwera, J. 2012. From Contrastive Linguistics to Linguistic Typology. *Languages in Contrast* 12(1): 69–86.

Viberg, Å. 1992. Universellt och språkspecifikt i det svenska ordförrådets organisation. *Tijdschrift voor Skandinavistiek* 13(2): 17–58.

Viberg, Å. 2006. Towards a Lexical Profile of the Swedish Verb Lexicon. *Sprachtypologie und Universalienforschung* (STUF) 59(1): 103–129.

Viberg, Å. 2013a. Seeing the Lexical Profile of Swedish through Multilingual Corpora. The Case of Swedish *åka* and other Vehicle Verbs. In *Advances in Corpus-based Contrastive Linguistics. Studies in Honour of Stig Johansson*, K. Aijmer and B. Altenberg (eds), 25–56. Amsterdam: Benjamins.

Viberg, Å. 2013b. Posture Verbs. A Multilingual Contrastive Study. *Languages in Contrast* 13(2): 139–169.

Viberg, Å. 2015a. Motion Verb Typology and the Expression of the Endpoint of Motion in Swedish. In *Concepts and Structures – Studies in Semantics and Morphology*. Studies in Linguistics and Methodology Vol. 8, M. Bloch-Trojnar A. Malicka-Kleparska and K. Drabikowska (eds), 209–229. Lublin: Wydawnictwo KUL.

Viberg, Å. 2015b. Contrasts in Construction and Semantic Composition: Crosslinguistic Perspectives on the Verbs of Putting in English and Swedish. In *Cross-linguistic Perspectives on Verb Constructions*, S.O. Ebeling and H. Hasselgård (eds), 222–253. Newcastle upon Tyne: Cambridge Scholars Publishing.

Viberg, Å. 2016a. What Happens in Translation? A Comparison of Original and Translated Texts Containing Verbs Meaning SIT, STAND and LIE in the English-Swedish Parallel Corpus (ESPC). *Nordic Journal of English Linguistics (NJES) 15*(3): 102–148. Available at: https://njes-journal.com/ [Last accessed 3 June 2021].

Viberg, Å. 2016b. Polysemy in Action: The Swedish Verb *slå* 'hit, strike, beat' in a Crosslinguistic Perspective. In *The Lexical Typology of Semantic Shifts*, P. Juvonen and M. Koptjevskaja-Tamm (eds), Cognitive Linguistics Research 58, 177–222. Berlin/Boston: de Gruyter Mouton.

Viberg, Å. 2017. Contrasts in Morphology. The Case of UP/DOWN and IN/OUT as Bound Morphemes in Swedish and their English Correspondents. In *Contrasting English and Other Languages through Corpora*, M. Janebova, E. Lapshinova-Koltunski and M. Martinková (eds), 32–74. Newcastle upon Tyne: Cambridge Scholars.

Viberg, Å. 2020. Contrasting Semantic Fields across Languages. In *New Approaches to Contrastive Linguistics. Empirical and Methodological Challenges*, R. Enghels, B. Defrancq and M. Jansegers (eds), 265–312 [Series: Trends in Linguistics. Studies and Monographs (TiLSM), 336]

Viberg, Å., Ballardini, K. and Stjärnlöf, S. 1984. *A Concise Swedish Grammar*. Stockholm: Natur & Kultur. For information about versions in other languages, see: https://www.nok.se/titlar/laromedel-b2/malgrammatiken/ [Last accessed 3 June 2021].

Author's address

Åke Viberg
Department of linguistics and philology
Uppsala University
Box 635
SE-751 26 Uppsala
Sweden
Ake.Viberg@lingfil.uu.se

Relativizers as markers of grammatical complexity: A diachronic, cross-register study of English and German[1]

Marie-Pauline Krielke

University of Saarland (Germany)

In this paper, we investigate grammatical complexity as a register feature of scientific English and German. Specifically, we carry out a diachronic comparison between general and scientific discourse in the two languages from the 17th to the 19th century, using relativizers as proxies for grammatical complexity. We ground our study in register theory (Halliday and Hasan, 1985), assuming that language use reflects contextual factors, which contribute to the formation of registers (Quirk *et al*., 1985; Biber *et al*., 1999; Teich *et al*., 2016). Our findings show a clear tendency towards grammatical simplification in scientific discourse in both languages with English spearheading the trend early on and German following later.

Keywords: contrastive linguistics, corpus linguistics, diachronic linguistics, English/German

1. Introduction

In the present paper, we look at the period between 1650 and 1900, which is especially interesting, since academic disciplines and with them scientific discourse emerges (Görlach, 2004). The development of new expressive structures reflects new communicative needs (cf. Betten, 2016). Register theory assumes that different text classes not only differ from general language in topic or field, but also in terms of lexico-grammatical features reflecting tenor and mode. This has been shown in numerous corpus-linguistic studies (Biber, 1988, 1993, 2006, 2012). Teich *et al.* (2016) follow the hypothesis that the development of scientific language undergoes two parallel processes, specialization and diversification. They show that, over time, scientific communication becomes increasingly expert-oriented, and the different scientific disciplines develop their own distinct set-ups of lexico-grammatical features. Specifically, for scientific English, previous research has shown a clear development towards higher lexical density (Biber, 2006; Aarts *et al*., 2012; Biber and Gray, 2016; Degaetano-Ortlieb *et al.*, 2016) alongside a simplification in syntax (Halliday, 1988; Teich *et al.*, 2016). German syntax, however, is described as becoming increasingly complex during the 17th and 18th century due to a strong remaining Latin influence and only in later periods, a trend towards detangling this

[1] This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 232722074 - SFB 1102.

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.v11i1.3440.

complex syntax is observed (Möslein, 1974; Beneš, 1981; Admoni, 1990; Habermann, 2011). Based on the findings for the two languages, we assume that decreasing grammatical complexity may be a cross-lingual register feature shaping scientific discourse over time.

We investigate grammatical complexity on the example of full, finite relative clauses (RCs). Being clausal postmodifiers, RCs represent the most explicit and syntactically most intricate way of defining a referent (as compared to alternative structures such as attributive adjectives and prepositional phrases) since at least a subject (specifying the agent) and a verb (marked for tense, aspect and mode) are included, as illustrated in example (1a). Transformations of RCs to less explicit structures, illustrated by the postmodifying prepositional phrase in example (1b), lead to processing difficulties characteristic of scientific language including (among others) lexical density, syntactic ambiguity and grammatical metaphor (cf. Halliday, 1988).

(1)

a) He affirms, that he has been the First *that* has discovered that Vessel, *which* by him is called Salivare Exterius. (Philosophical Transactions, 1665−1678)

b) He affirms his discovery *of the Salivare Exterius Vessel*. (generated alternative)

Due to clausal embedding example (1a) is grammatically more complex than example (1b). Accumulations of RCs within one sentence represent especially strong cases of syntactic complexity, as illustrated in example (2).

(2) Next, that the two Eyes were united into one Double Eye, *which* was placed just in the middle of the Brow, the Nose being wanting, *which* should have separated them, *whereby* the two Eye-holes in the Scull were united into one very large round hole, into the midst of *which*, from the Brain, entered one pretty large Optic Nerve, at the end of *which* grew a great Double Eye; that is, that Membrane, called Sclerotis, *which* contained both, was one and the same, but seemed to have a Seam, by *which* they were joined, to go quite round it, and the fore or pellucid part was distinctly separated into two Cornea's by a white Seam *that* divided them. (Philosophical Transactions, 1665−1678)

Besides the most common relativizers *which* and *that*, pronominal adverbs are another, highly explicit way of conveying a relationship between the antecedent and the subject of the RC, as in example (3a).

(3)

a) […] the Membrane immediately encompassing that skin, *wherein* the Faetus is wrapped […]. (Philosophical Transactions, 1665−1678)

b) […] the Membrane immediately encompassing that skin wrapping the Faetus […]. (generated alternative)

As seen in example (3a), the prepositional specification of location *in* (in where*in*) is implicitly entailed in the verb *wrap* and could therefore be omitted (3b). Omission of any kind of superfluous grammatical information could be a counterbalance to other sources of informational overload, such as continuously new emerging vocabulary. Besides lower syntactic intricacy, a reduced set of alternatives at a given choice point, i.e., fewer different relativizers and better predictability of the specific options reduce grammatical complexity through reduction of entropy.

2. Related work

Relative clauses (RCs) are a widely studied topic in English diachronic as well as synchronic linguistic studies (Ball, 1996; Nevalainen and Raumolin-Brunberg, 2002; Hundt *et al.*, 2012; Nevalainen, 2012), in vernacular varieties of English (Romaine, 1980, 1982; Tottie and Harvie, 2000; Tagliamonte, 2002; Tagliamonte *et al.*, 2005; Levey, 2006) and in spoken and written mode (Guy and Bayley, 1995; Lehmann, 2001). Diachronic as well as synchronic studies (Biber *et al.*, 1999; Leech *et al.*, 2009; Hinrichs *et al.*, 2015) on relativizer choice find that the selection of relativizers largely depends on the overall formality level of a text, *which* being the formal option whereas *that* becomes increasingly common to informal text types. However, there are only few studies reflecting on the use of pronominal adverbs in relativizer position. Mellinkoff (2004), for instance, mentions their diachronic integration in the language of the law, and Österman (1997) points to their primary association with formal genres. Diachronically, Krielke *et al.* (2019) have shown a remarkable decrease in pronominal adverbs in scientific English between 1650 and 1850. This paradigm reduction of pronominal adverbs over time can partly be explained by the typological drift from synthetic to analytic (Nevalainen and Raumolin-Brunberg, 2012), e.g., *whereby* becoming *by which*.

RCs and their and alternative syntactic renderings (prepositional phrases and attributive adjectives), as well, have received ample scholarly attention for their role as frequent constituents in noun phrases. For written discourse, Biber *et al.* (1988) report on a strong preference for (premodified) nouns and postmodifying prepositional phrases, while spoken registers rather rely on embedded clauses. Biber and Finegan (1997) show a steady trend towards nominal structures in the past 300 years of academic writing. Biber and Gray (2011) specifically mention a slight decrease in RCs in academic texts, again pointing to a remarkable increase in phrasal as compared to clausal modification creating a compressed academic style in present day English (see e.g., Halliday, 1988; Biber and Clark, 2002; Mair, 2006; Biber and Conrad, 2009).

The aforementioned studies largely rely on patterns of parts-of-speech, but also syntax-based studies found a decrease in RCs as compared to nominal premodifications (see for instance Juzek *et al.*, 2020). The predominantly frequency-based approaches, however, ignore ambient context. Information-theoretic measures, such as surprisal and entropy, considering the probabilities of linguistic units given their syntagmatic contexts have proven to be important factors driving linguistic change (Degaetano-Ortlieb and Teich, 2016; 2019; Rubino *et al.*, 2016). Especially convergence on specific grammatical features, which can be measured by conditioned probabilities, over time leads to conventionalization, a mechanism giving way to innovation on the lexical level. Degaetano-Ortlieb and Teich (2019) give a unified explanation of the evolution of the scientific register based on the assumption that register evolution depends on evolving communicative needs while striving for the creation of an optimal code customized for communication between experts. This code is assumed to be characterized by specific linguistic features to balance information load. For our study we adopt Teich *et al.*'s (2016) assumptions regarding register shifts formulated in their hypotheses on specialization: The development of a scientific field leads to increasing expert-orientation. Expert-orientation manifests itself along two dimensions: a) increasing technicality and information density, linguistically expressed by nominal style and high lexical density, and b) decreasing grammatical intricacy of the sentence structure, i.e., the number of clauses in the sentence and their interdependencies (cf. Halliday, 1988; Halliday and Martin, 1993). To measure grammatical intricacy, Teich *et al.* (2016) inspect, amongst other features, the number of clauses (including RCs) as well as the number of relativizers per sentence in scientific texts between the 1970s and the 2000s. For our study this points to the assumption that RCs are a

feature of scientific discourse, however intricacy in the sense of embeddedness of many clauses within one sentence is rather specific to general language.

In contrast to the studies of English, studies of German on diachronic grammatical change are more qualitative in nature. The observed time period in this study starts at the end of the third and last period of Early New High German, a time period bringing forth a variety of new, especially informative text types promoting increasing distinctiveness between general language and the language of the learned (1550−1700; Admoni, 1990).

Habermann's (2001) comprehensive account on the development of German syntax in the natural sciences between the 15th and 19th century focuses on the influence of Latin on the emerging vernacular German as a language of scientific communication. Scientists received their education in Latin, influencing their lexical as well as syntactic style. Preferred structures influenced by Latin were, for instance, sentence equivalent short forms pursuing information density, while expanding hypotaxis with deep embeddings was preferred over parataxis.

Möslein (1974) describes the syntactic developments in scientific-technical literature since the end of the 18th century. Due to the establishment of verb final position in the 17th century (starting in technical literature), main and subclause can be distinguished from each other. Formation of long and embedded sentences to present complex thoughts in one sentence becomes possible leading to an extreme increase in hypotactic structures in the 17th and 18th century (ibid.). Starting in the first half of the 19th century, a trend of disentanglement and reduction in sentence length as well as a remarkable reduction in subordinate clauses and an increase in nominalizations is described to take place in scholarly German (Möslein, 1974; Beneš, 1981). Societal developments of the time, such as increasing influence of mass media and other European languages of science, are reflected in a new trend towards lower syntactic intricacy. As a result, scientific style became increasingly condensed aiming for clarity and efficiency of expression in response to evolving communicative needs. Possible reasons for the increase of nominal groups instead of subclauses could be exactness and effort reduction (see for instance *dependency locality theory*, Gibson *et al.*, 2000). Factors that may lead to reduction in cognitive effort are compound formation as an alternative to prepositional phrases (*iron oxide vs. oxide of iron*), and nominalizations instead of subclauses avoiding grammatical complexity connected to tense, mode and number (Möslein, 1974).

The studies on the different German relativizers we are aware of (Ebert, 1986; Reichmann and Wegera, 1993; von Polenz, 1999; Ágel, 2000; Fleischer, 2004; Brooks, 2006; Dal, 2014; Pickl, 2020) only look at the standard relativizers, *der/die/das (d-)* being the most frequent relativizer and *welcher/welche/welches (welch-)* being the marked, formal variant (see Pickl, 2020) in isolation, while a comprehensive view on relativizers, including pronominal adverbs, is still lacking. Mentioned as promoters of syntactic intricacy (Möslein, 1974, Admoni, 1990), RCs have also been analyzed with information-theoretic measures. Voigtmann and Speyer (2020), for instance, use surprisal (Shannon, 1949; Levy, 2008; see section 4.2) to detect information density related preference for RC extraposition, assuming that extraposition is used as a strategy to counterbalance an informational overload in the RC and spread information evenly across a sentence. Krielke *et al.* (2019) use surprisal to detect increasingly predictable contexts of the relativizer *which* in English, finding that it is increasingly used in adverbial gaps (see Biber *et al.*, 1999), particularly to express relations of manner (*the means by which, the manner in which*). Further, Krielke *et al.* (2019) use entropy (see section 4.2) to trace paradigmatic variety of a group of relativizers including prepositional adverbs similar to Milin *et al.* (2009), who measure entropy over inflectional paradigms. In the present study we use entropy to measure paradigmatic variability of the relativizer paradigm, as well as surprisal to account for the predictability of relativizers in German and English over time. We assume that in scientific discourse the contexts of RCs become increasingly predictable over time, thus contributing to the overall processing ease of otherwise complex concepts.

3. Hypotheses

We use relativizers as proxies to investigate the development of grammatical complexity for English and German, pursuing the following hypotheses:

1. In the course of register formation, scientific discourse becomes grammatically less complex in terms of

   a) *syntactic intricacy*, as indicated by the frequency of relativizers and the number of RC embeddings within a sentence, decreases as register formation evolves.

   b) *paradigmatic richness*, i.e., the available types of relativizers, decreases over time as indexed by entropy.

   c) *contextual predictability* of relativizers, i.e., relativizers appear in increasingly similar contexts as indexed by surprisal.

Since German is described to initially expand its syntactic possibilities during the 18th century and due to much later institutionalization of scientific discourse in German, we expect the development to manifest along different trajectories, leading to the second hypothesis:

2. English should show a more linear development, while we expect German to first increase syntactic intricacy and paradigmatic richness before decreasing towards the 19th century.

4. Data and Methods

4.1 Data

For scientific English (SE), we use the *Royal Society Corpus* (RSC v4.0; Kermes *et al.*, 2016), consisting of the *Proceedings and Transactions of the Royal Society of London* covering the time from 1665−1869 with approximately 32 million tokens. For general English (GE), we use the *Corpus of Late Modern English Texts* (CLMET; Diller *et al.*, 2011), spanning 1710−1920 with approximately 40 million tokens from several genres (e.g., narrative, drama). For German, texts from 1650−1900 are retrieved from the scientific (SG) and general language (GG) subcorpora of *Deutsches Textarchiv* (DTA, Geyken *et al.*, 2018) respectively. Scientific German is represented with approximately 80 million tokens, general German with approximately 60 million tokens including non-fictional as well as fictional prose texts. All subcorpora contain metadata (e.g., author, publication year) and linguistic annotation (e.g., tokens, lemmas, normalization, parts-of-speech, surprisal). Part-of-speech annotation for English is based on the "Penn Treebank Tagset" (Santorini, 1990), for German on the "Stuttgart-Tübingen Tagset" (STTS, Thielen *et al.*, 1999).

4.2 Methods

To trace the development of grammatical complexity in the scientific register in the two languages, we focus on the features shown in Table 1. We apply conventional frequency-based methods to account for syntactic intricacy indicated by the frequency of relativizers in the four subcorpora, as well as the number of RC embeddings within a sentence.

Table 1. Features of grammatical complexity.

Discourse Property	Feature Category	Feature subcategory	Measure
Grammatical complexity	Grammatical Intricacy	Frequency of relativizers Relativizers per sentence	Relative frequencies
	Paradigmatic richness	Relativizer paradigm (number of different relativizers)	Entropy (H)
	Contextual predictability	Probability of relativizers given their context	Surprisal

To account for predictability of items in context, we use information-theoretic measures such as surprisal (as operationalized by Degaetano-Ortlieb *et al.*, 2016) of the different relativizers. To assess paradigmatic changes (growth or reduction) in the group of available relativizers, we calculate entropy. Finally, we qualitatively investigate the top three preceding trigrams sequences (part-of-speech and lexical) representing highly predictive contexts of relativizers.

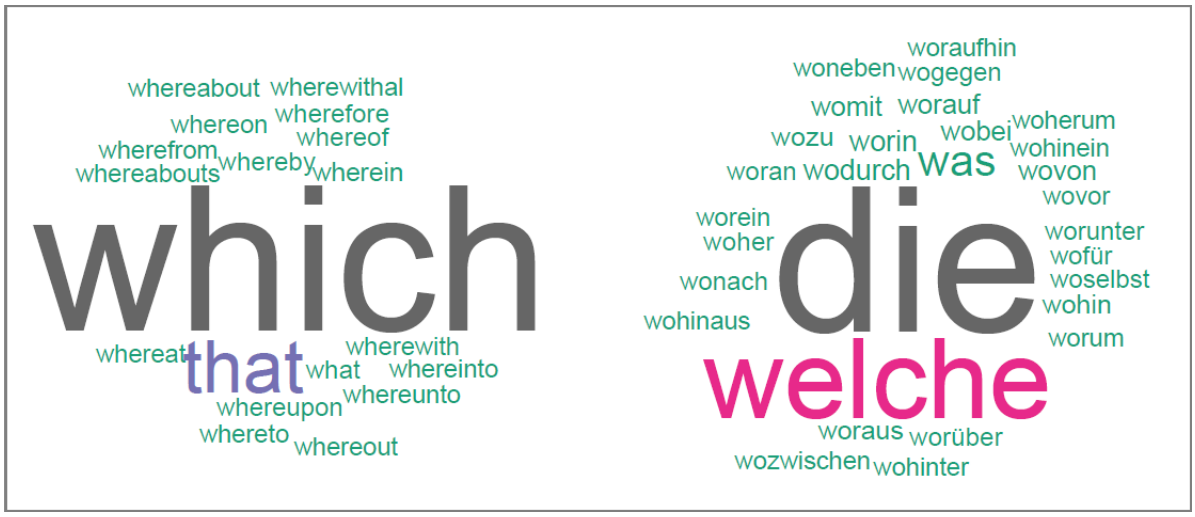

Figure 1. Relativizer paradigms in English (left) and German (right), color and size indicate frequency.

To grasp the full historical extent of the paradigms (apart from the most common ones *which* and *that* and *welcher, welche, welches (welch-)* and *der, die, das (d-)*), we first determine the existing members of the paradigms. For English, we extract all words beginning with *where-* and sort out all words not representing pronominal adverbs and, for German, all words beginning with *wo(r)-* and part-of-speech (POS) tagged as PRELS/PRELAT, resulting in the lists provided in figure 1. The motivation to use information-theoretic measures is the assumption that language users strive for effort reduction on the one hand and successful communication on the other. Previous studies have shown that production effort is directly linked to the number of options at a given choice point (Milin *et al.*, 2009). Fewer encoding options lead to entropy reduction (cf. hypothesis 1b). For the relativizer paradigm we can assume that fewer available relativizers lead to lower production effort for the sender of a message, as well as lower comprehension effort for the receiver of the message, since that receiver will have a more confined expectation and lower uncertainty of the upcoming word. We use entropy to measure the uncertainty about a set of choices at a given point. Formally, entropy is the expected (weighted average) amount of information in a paradigm. The more members the paradigm has and the more similar the probabilities of the different members are, the higher the entropy. Thus, entropy is highest if all probabilities are equal. We calculate the entropy of the English and German relativizer paradigms (figure 1) to find whether there is a register specific trend for entropy reduction and if so, whether this is the case in both languages.

Register specific preference and with it a reduction in paradigmatic entropy should lead to convergence on conventionalized linguistic choices.

Entropy is directly related to surprisal, i.e., the negative *log* probability of a word to occur in a certain context (Crocker *et al.*, 2015). The higher the probability of a word in a particular context, the less surprising is its occurrence in this context (Degaetano-Ortlieb *et al.*, 2016). We calculate surprisal based on the conditional negative *log* probabilities from a 4-gram language model, i.e., the negative log probability of a word given its three preceding words. For our analysis, we are interested in the distributions of the surprisal values of the observed three groups of relativizers (*which, that, (welch-), (d-)* and pronominal adverbs). To visualize the surprisal distributions of each relativizer group, we use boxplots displaying the distribution of the individual surprisal values indicating five different measures: minimum, first quartile, median, third quartile and maximum. The boxplots also show whether the values are symmetrically distributed, how tightly the values are grouped and if they are skewed. We assume that, over time, relativizers will occur in increasingly predictable contexts (have a lower surprisal), ensuring successful and effortless communication (cf. hypothesis 1c).

5. Analysis

5.1 Syntactic intricacy

We aim to trace changes in syntactic intricacy via two measures. First, we calculate (a) relative frequencies of the whole group of relativizers per subcorpus per 50 years, assuming that a higher number of relativizers represents a higher number of RCs and therefore a preference for post-modification. Second, we calculate (b) the average number of relativizers per sentence per 50 years, assuming that a higher number of relativizers per sentence represents a stronger tendency towards embeddedness. (a) is displayed in figure 2 (for English) and 3 (for German).

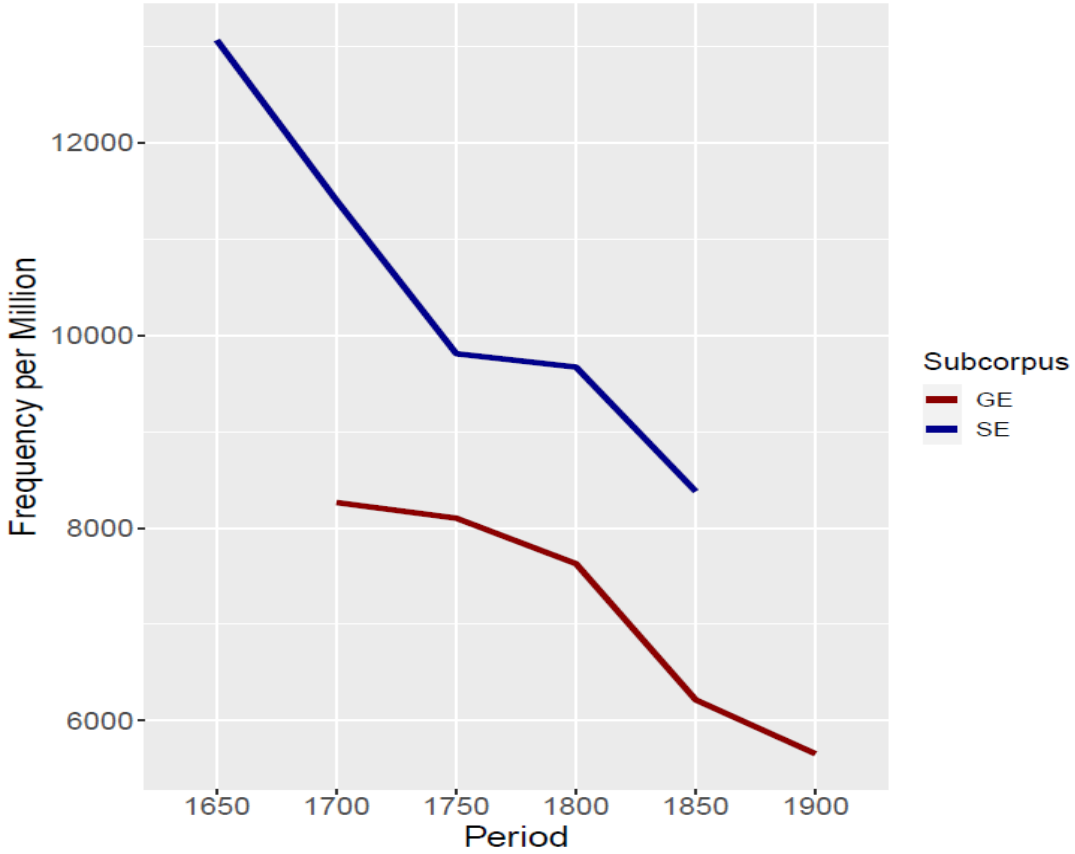

Figure 2. Frequency per million of relativizers in general (GE) vs. scientific English (SE).

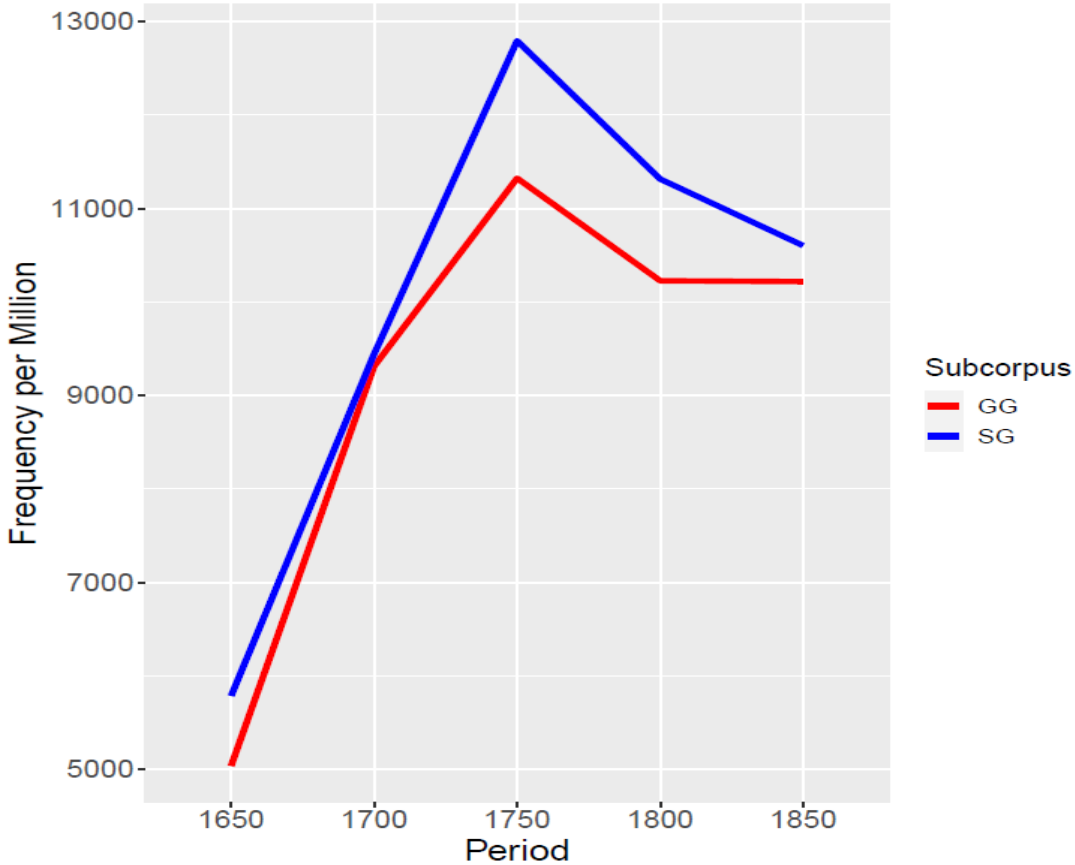

Figure 3. Frequency per million of relativizers in general (GG) vs. scientific German (SG).

First, between languages, we see strongly diverging trends, while within languages trends are quite similar. For English, we see a steady decrease of relativizers in both scientific and general language. In the scientific texts, relative frequencies start out almost twice as high and with a much steeper downward trend than in the GE texts. This shows an overall preference for using RCs in SE in the earlier time periods and a clear development towards a less embedded syntax over time. In German, the frequencies in both subcorpora are quite similar until 1750 and start to diverge afterwards. Also, throughout the observed time period SG shows higher frequencies of relativizers than GG. Frequencies peak in 1750 followed by a decrease in both SG and GG. The peak in SG, however, is more pronounced. In the first half of the 19th century frequencies stabilize in GG, while further declining in SG. In contrast to English, German starts out at a much lower frequency of relativizers between 1600 and 1650 only reaching the starting frequency of SE in 1750, indicating that in German hypotaxis was expanding throughout the baroque period, as also suggested by Habermann (2001) and Admoni (1990). The trends differ between the two languages until 1750 and align afterwards indicating that in German the trend towards simpler syntax became popular only in the 18th century as suggested by Admoni (1990). This is comprehensible considering the strong influence of Latin stylistic ideals German scientific text production was under.

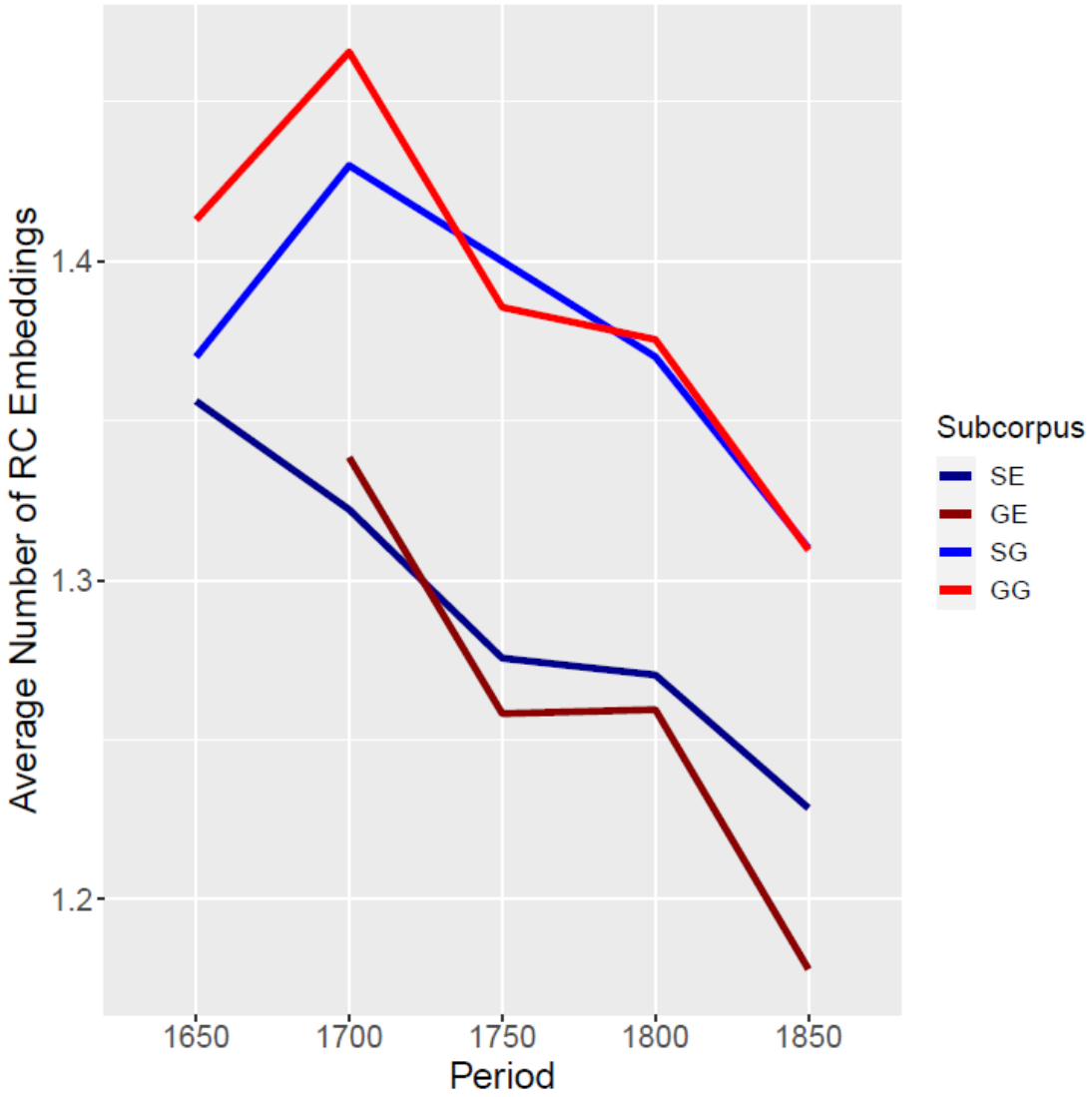

Figure 4. Average number of RCs per sentence in English (GE & SE) and German (GG & SG).

Looking at the average number of RCs embedded in a sentence (figure 4), we find that the trends broadly coincide with the shapes of the frequency distributions. In the first half of the 18th century embeddedness is stronger in GE than in SE, while there are overall more relativizers used in SE than in GE. This indicates that GE overall made use of fewer RCs, which, however, often occurred within one sentence. In the second half of the 18th century the trend reverses. Scientific English shows stronger embeddedness together with a higher number of RCs overall. In both SG and GG, RC embeddings show a steep increase towards the first half of the 18th century and an equally steep decrease afterwards representing the flourishment of clause embeddings in the 17th and 18th century. Interestingly, SG overall shows fewer embeddings per sentence than GG (with an exception between 1750 and 1800) while constantly showing a higher frequency in relativizers. This points to a need to employ explicit structures to explain complex matters while not overstretching the boundaries of cognitive processing load. Another interesting fact is that RC embeddedness peaks earlier than the overall frequency of relativizers. This points to a rather unbalanced use of relativizers in the first half of the 18th century: fewer relativizers overall clustering together in fewer sentences. In the second half of the 18th century this trend reverses: more relativizers overall are spread across different sentences.

5.2 Paradigmatic richness

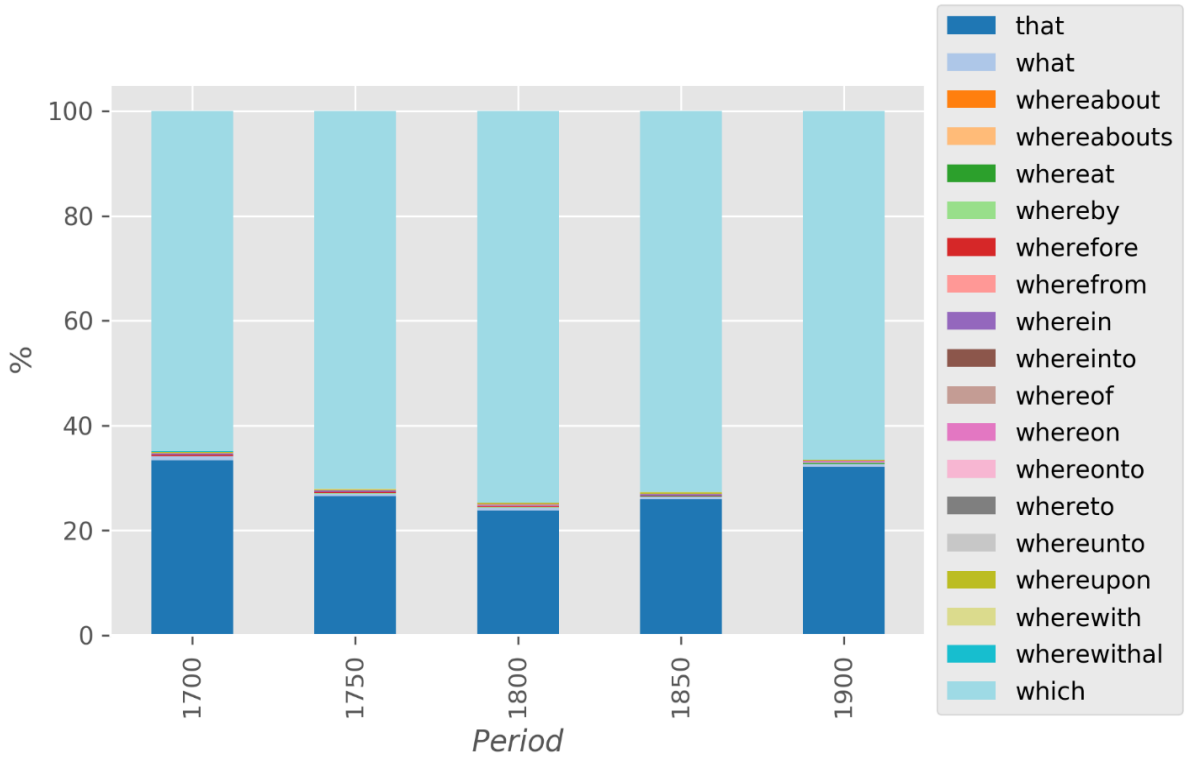

Figure 5. Distribution of different relativizers in GE.

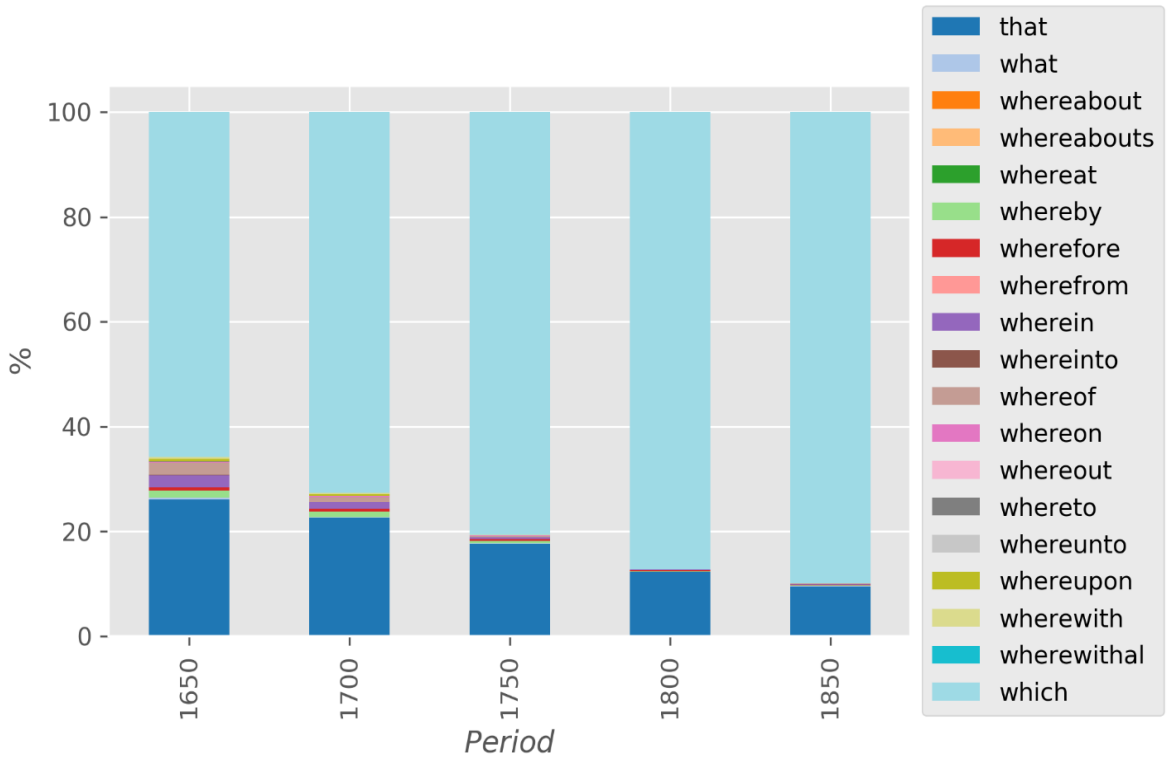

Figure 6. Distribution of different relativizers in SE.

The English subcorpora differ substantially regarding relativizer distribution, making relativizers a clear register feature in the early periods. While GE (figure 5) shows a stable distribution of the different relativizers with *which* being the overall most frequent type, SE (figure 6) starts out with a great variability of available relativizers including a large group of pronominal adverbs. For the GE subcorpus, we see that pronominal adverbs throughout all time periods occupy a rather negligible proportion. Looking at the distribution of the different relativizer types in SE, we find the biggest variability of relativizers in 1650. Together with overall decreasing relative frequencies of relativizers, variability gradually decreases, too. Over time, *which* becomes increasingly dominant, pushing out all other relativizers to under 10% in 1850. The gradual decrease of pronominal adverbs is in line with observations by Nevalainen and Raumolin-Brunberg (2012) and Krielke *et al.* (2019), confirming the abandonment of synthetic forms. In addition, this outcome confirms our intuition about differentiation of the scientific register against the general language by converging on a preferred linguistic feature and substituting a variety of alternatives. This is in line with observations of conventionalization in the scientific domain by Degaetano-Ortlieb and Teich (2019) and Teich *et al.* (2021). We will show this even more clearly by calculating entropy of the relativizer paradigms in each subcorpus.

In German, we find an inverse picture. Figure 7 shows that, like SE, in GG pronominal adverbs become less frequent. In SG (figure 8), in contrast, pronominal adverbs take up an increasing portion of the paradigm until 1850 and abruptly decrease after 1850. In exchange, *(welche-)* becomes more frequent taking the place left by the pronominal adverbs in decline. This is interesting for two reasons. First, German academic style seems to prefer a diverse set of options to introduce RCs in an explicit way. Only towards the end of the 19th century both frequency and relativizer variety seem to decline, possibly following the example of other European scientific traditions as suggested by Möslein (1974) and Beneš (1981). Second, while overall relativizer frequency was already on the wane, productivity of relativizers was still in

expansion: This points to an even greater variability of the paradigm in the period between 1800 and 1850.

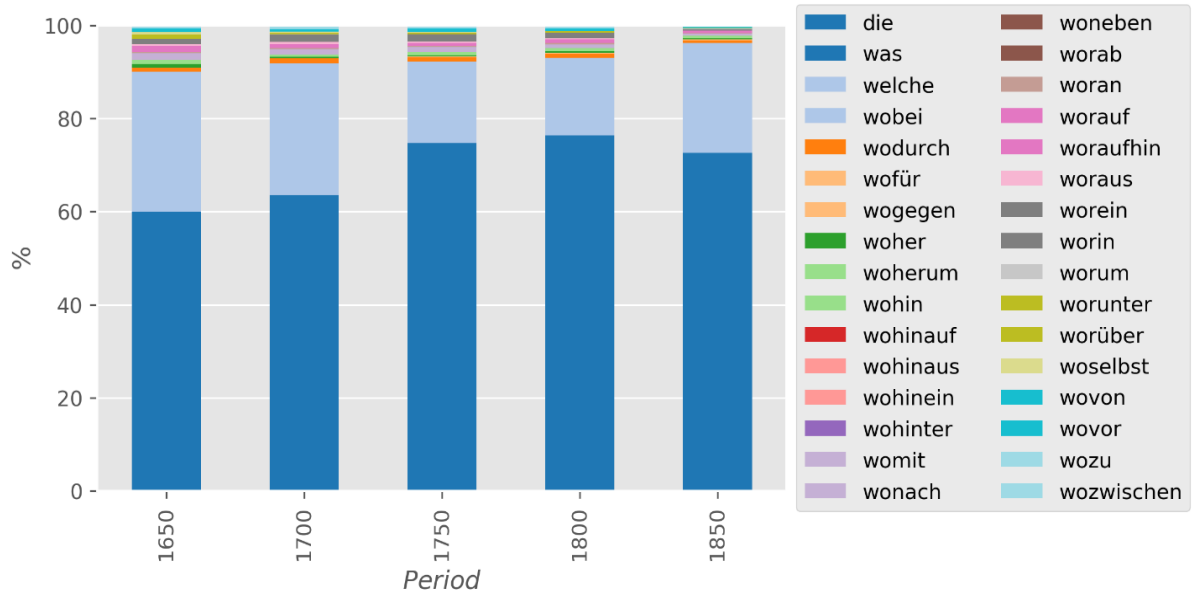

Figure 7. Distribution of different relativizers in GG.

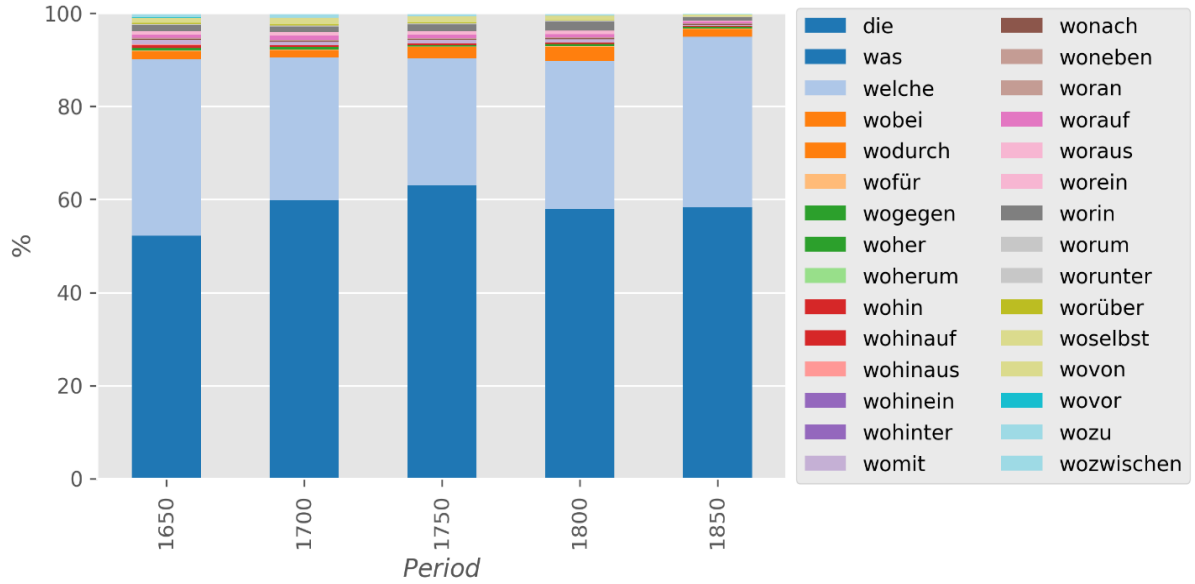

Figure 8. Distribution of different relativizers in SG.

Looking at entropy, we find relatively stable values in GE (figure 9), while for SE (figure 10) we see a striking reduction of entropy over time. The entropy trends clearly reflect the distributional trends in figures 5–6, while also considering predictability: the reduction in entropy in SE over time is owed to a smaller choice of options between the different relativizers on the one hand, but also to an increasing probability of *which* to occur as compared to decreasing probabilities of all other available options.

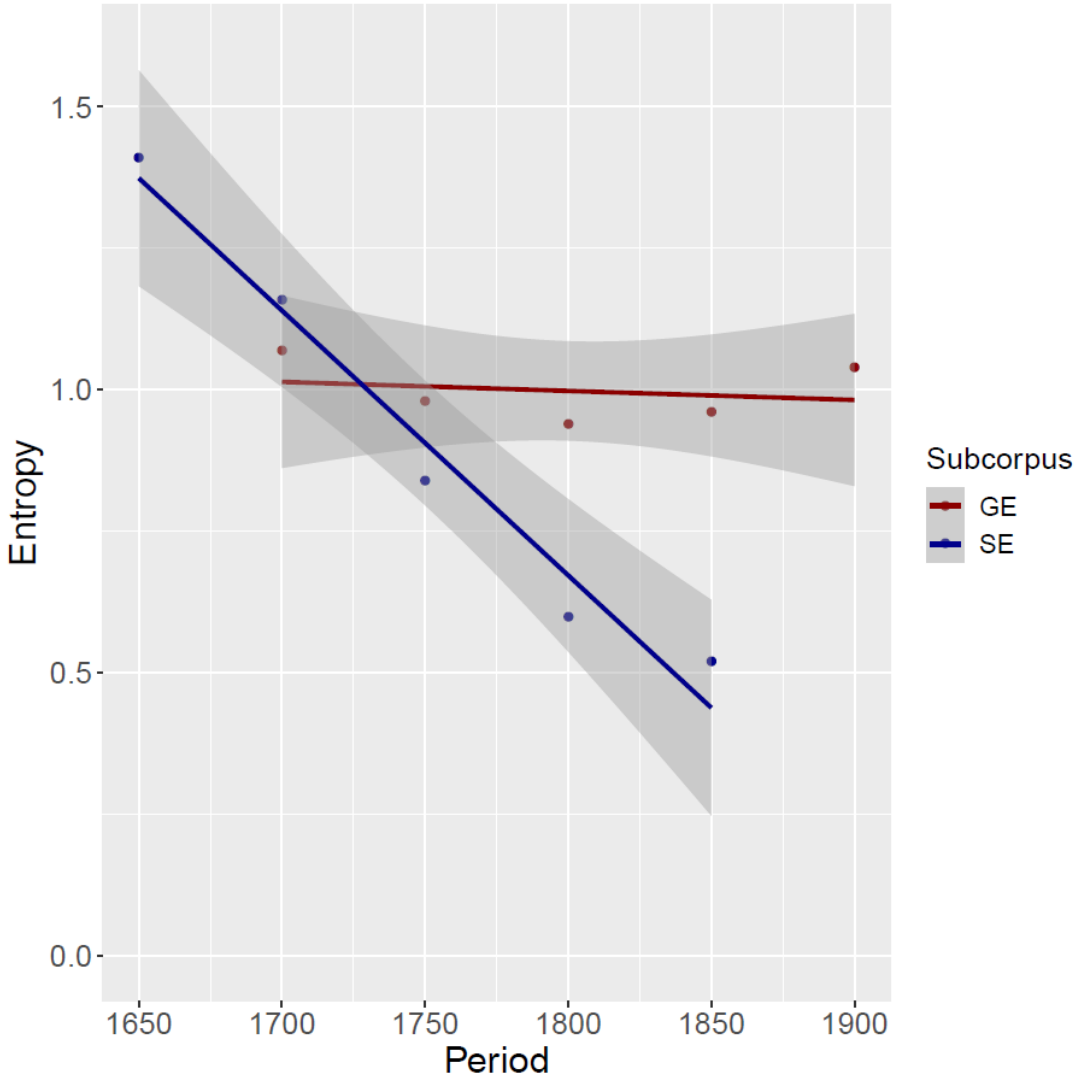

Figure 9. Entropy over the relativizer paradigm in general (GE) and scientific English (SE).

Figure 6 shows that in 1650 scientific writers had a much bigger choice amongst different relativizers than in 1850. At the same time, readers of scientific texts in 1650 had a much higher uncertainty about the upcoming relativizer than a reader in 1850. In GE (figure 5), the choice/uncertainty did not change over time. The entropy value of 1 points at a choice between two preferred options, presumably *that* and *which*. The entropy value in SE in 1850 is around 0.5, a third of the value in 1650, indicating a strong preference for *which* as the relativizer of choice. This again reflects the tendency towards conventionalization of options, as observed by Teich *et al.* (2021). In German, the paradigm of relativizers is much bigger than in English (compare figure 1), contributing to overall higher entropy values (German ranges between 1.5 and 2, while English ranges between 0.5 and 1.5).

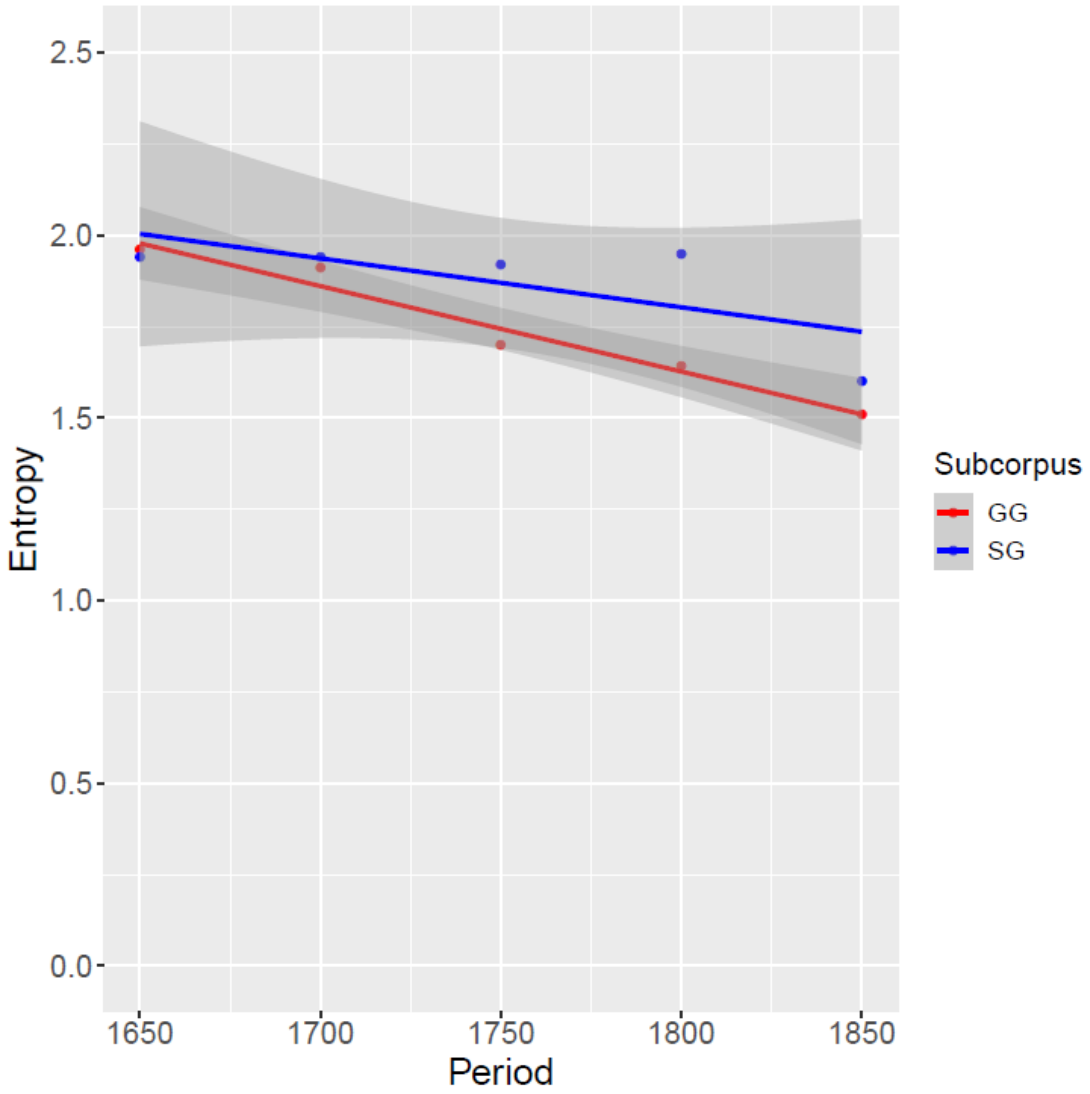

Figure 10. Entropy over the relativizer paradigm in general (GG) vs. scientific German (SG).

Figure 10 shows that entropy in GG steadily decreases after 1650, while in SG entropy is relatively stable (at approx. 1.9) until 1850 and falls after that. In 1850, entropy in SG is almost as low as in GG. During the period between 1650 and 1800, SG seems to prefer a richer choice of possible options over a monopoly of few options, whereas GG continuously develops towards a more confined set of options. Consistent with the rise in frequency of relativizers in SG until 1850, entropy, too, reflects increasing complexity regarding relativizer use and a drop thereof afterwards. For English, the results of our entropy calculations show a clear distinction between SE and GE, pointing to a clear development of a register specific preference of *which* in scientific language. In general language, however, the choice seems to be between *which* and *that*. In German, the stronger tendency of scientific texts towards diversity in relativizer choice during the period between 1650 and 1850 confirms Admoni's (1990) and Habermann's (2004) observation of expanding grammatical complexity in the scientific genre. The final drop in entropy towards 1900 reflects an eventual turn towards fewer options at the choice point of the relativizer.

5.3 Contextual predictability

We calculate surprisal for the different groups of relativizers (*which/(welch-), that/(d-)* and pronominal adverbs in order to see whether they become more or less predictable in context over time.

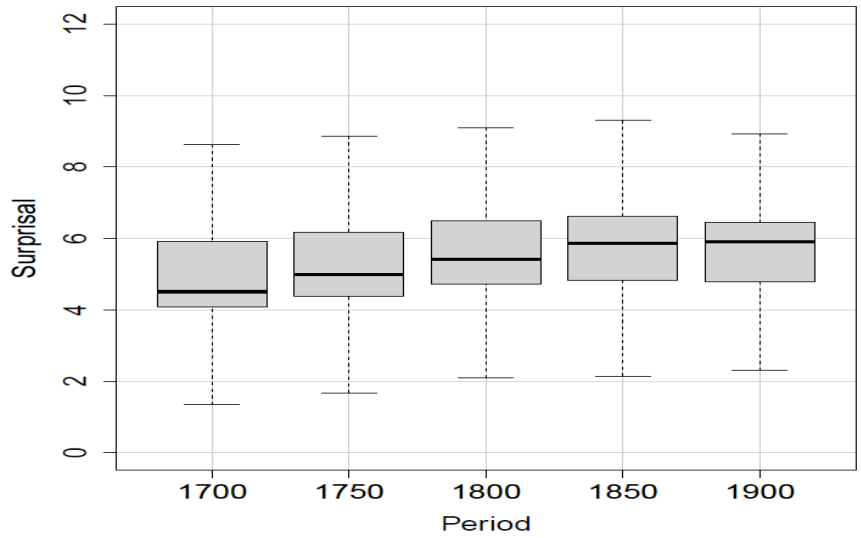

Figure 11. Distribution of surprisal values for "that" per 50 years in GE.

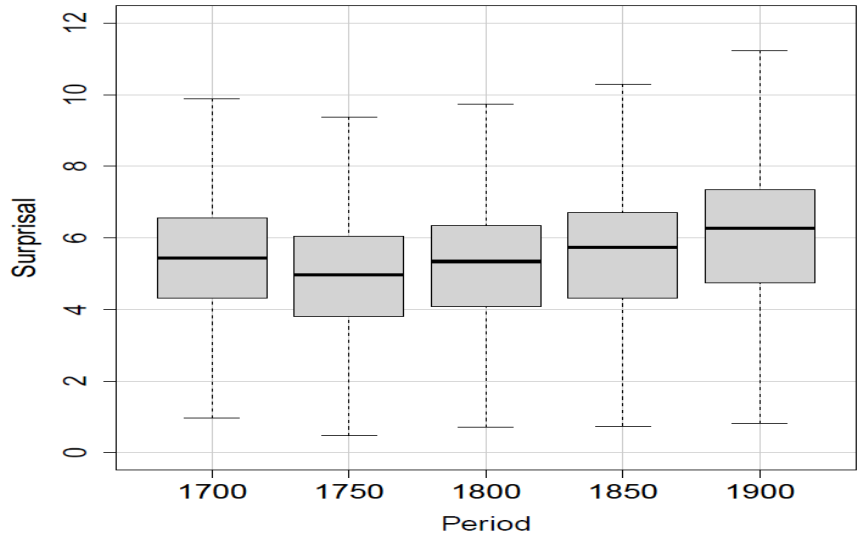

Figure 12. Distribution of surprisal values for "which" per 50 years in GE.

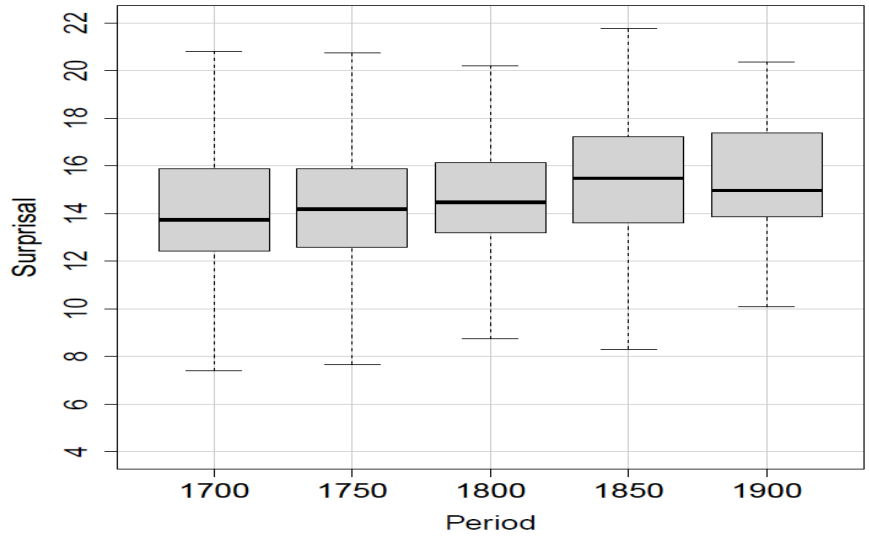

Figure 13. Distribution of surprisal values for pronominal adverbs per 50 years in GE.

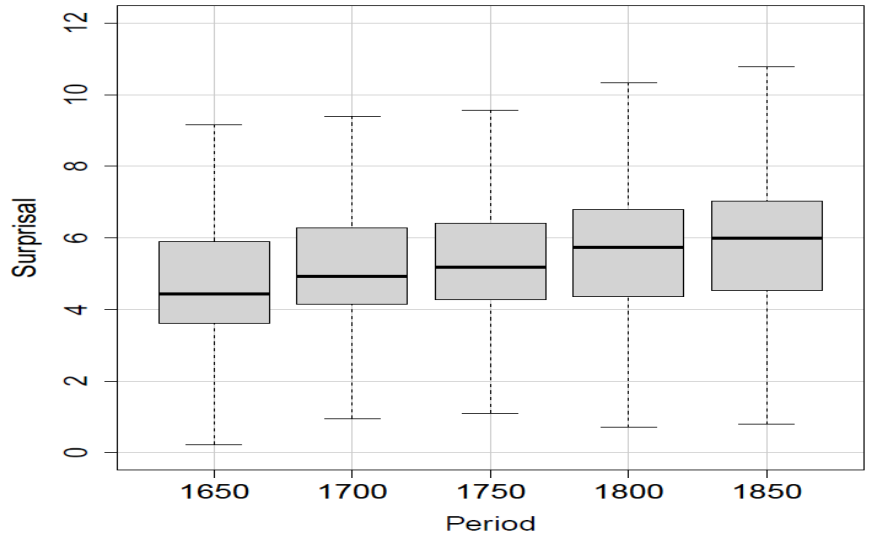

Figure 14. Distribution of surprisal values for "that" per 50 years in SE.

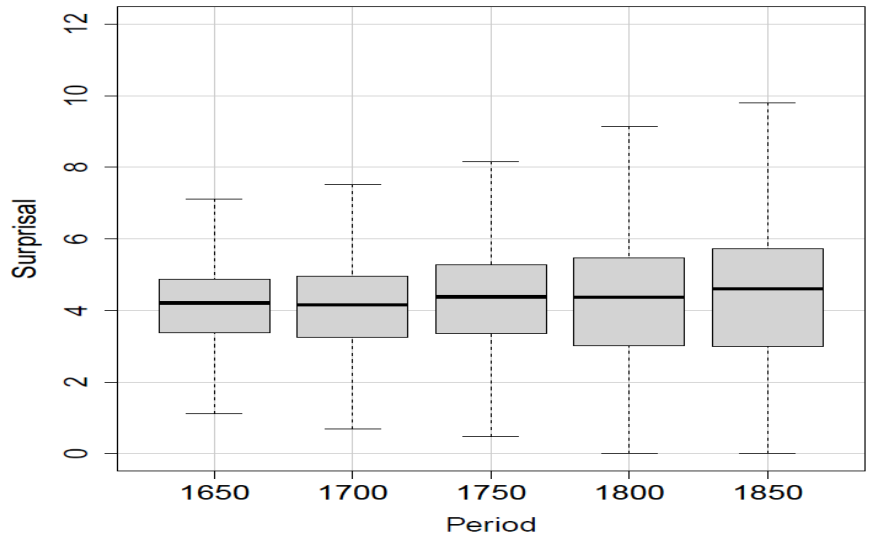

Figure 15. Distribution of surprisal values for "which" per 50 years in SE.

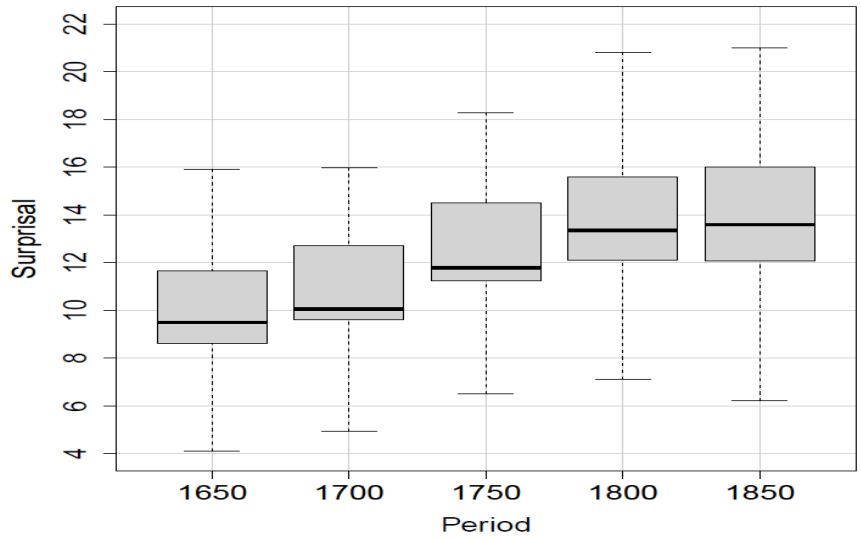

Figure 16. Distribution of surprisal values for pronominal adverbs per 50 years in SE.

The English subcorpora (figures 11−13 for GE, figures 14−16 for SE) are fairly similar for all three relativizer types. All of them become more surprising over time reflecting the decrease in use. *That* (figures 11 and 14) shows a slight increase in median surprisal values from about four to six. In GE, the maximum and minimum surprisal values (whiskers) diverge less than in SE indicating that *that* becomes more stable in terms of predictability in GE compared to *that* in SE. This is plausible since *that* is less frequently used in the latter. The surprisal values for *which* in GE (figure 12) are slightly higher than in SE (figure 15), reflecting the lower frequency in general language. In both subcorpora, the range of surprisal values increases over time. The long whiskers indicate a broader use of *which* in 1850 compared to 1650 with very frequent patterns of usage (very low surprisal values) as well as very infrequent ones (very high surprisal values). This tendency is especially evident in SE – plausibly so, since *which* over time becomes the number one relativizer in SE, while all other relativizers become less frequent. Thus, the contexts formerly covered by different relativizers are now filled by *which*. At the same time, *which* shows the most stable median surprisal values over time. This indicates that most of its preferred contexts are stable with some of them becoming particularly conventionalized and thus unsurprising. This development is in line with the theory of conventionalization in scientific language put forward by Degaetano-Ortlieb and Teich (2019) and Teich *et al.* (2021). Surprisal of pronominal adverbs in GE (figure 13) shows the constantly highest surprisal. Surprisal in SE (figure 16), instead, starts out much lower increasing continuously as pronominal adverbs and with them the possible contexts become continuously less frequent.

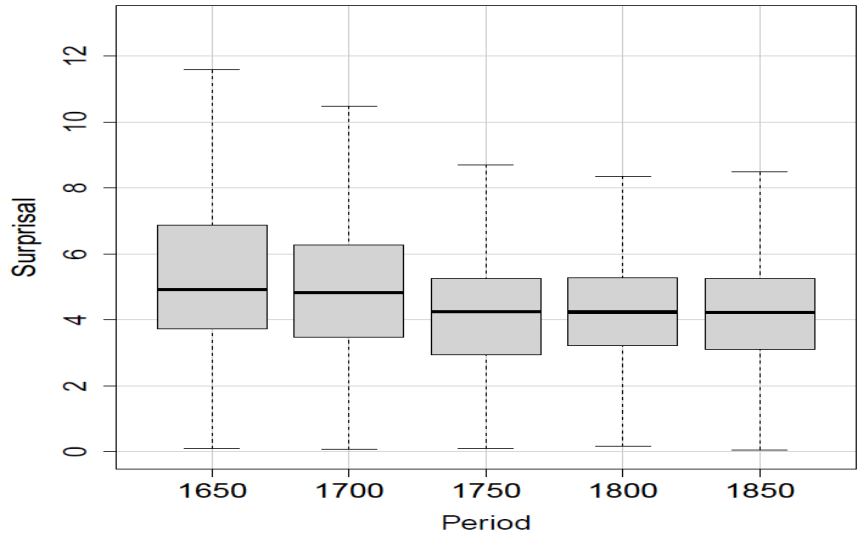

Figure 17. Distribution of surprisal values for (*d*-) per 50 years in GG.

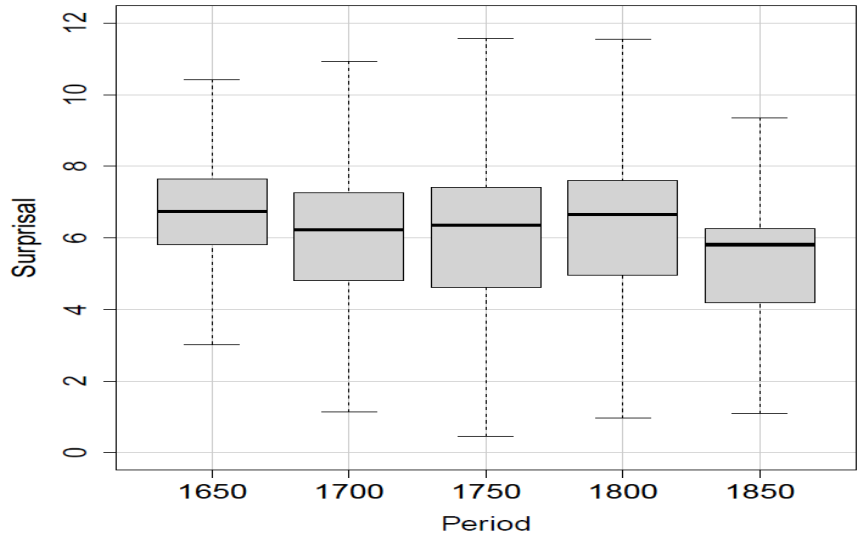

Figure 18. Distribution of surprisal values for (*welch*-) per 50 years in GG.

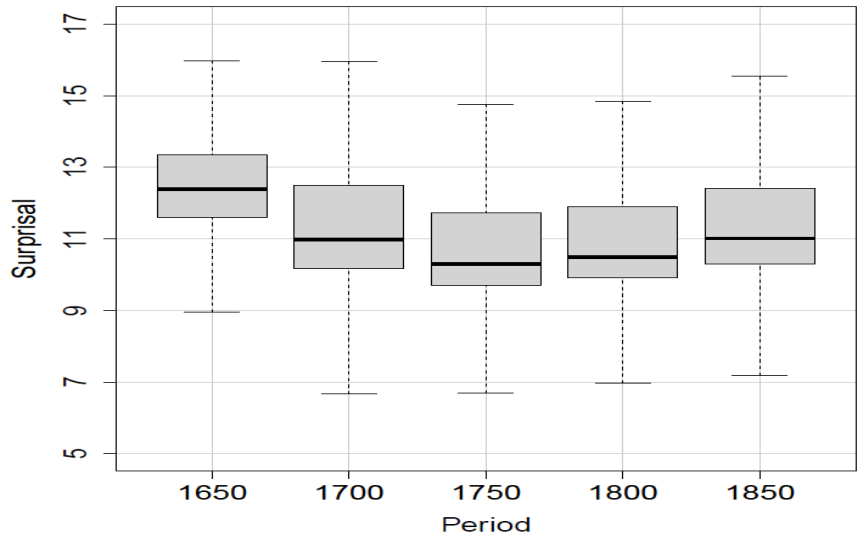

Figure 19. Distribution of surprisal values for pronominal adverbs per 50 years in GG.

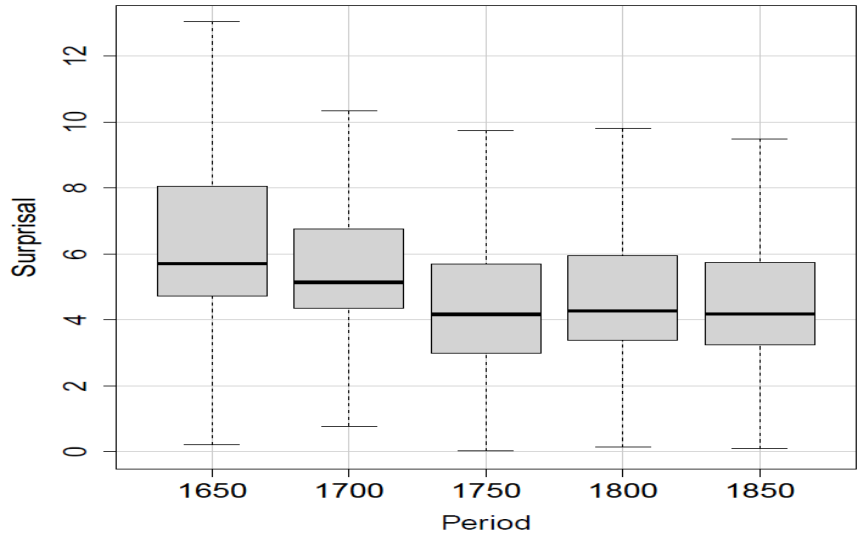

Figure 20. Distribution of surprisal values for (*d*-) per 50 years in SG.

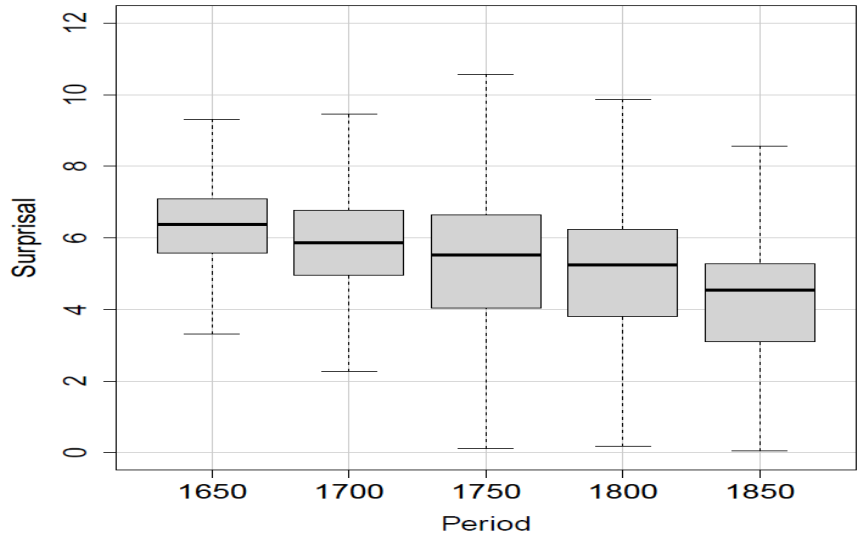

Figure 21. Distribution of surprisal values for (*welch*-) per 50 years in SG.

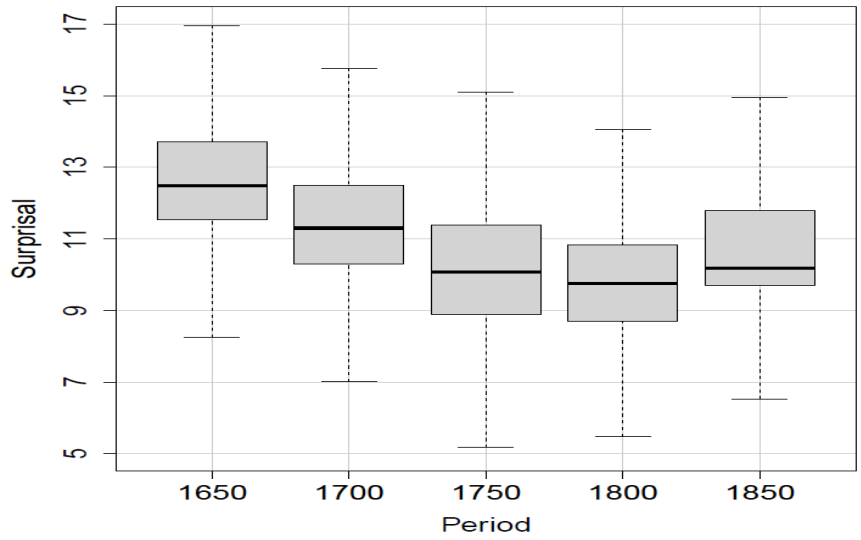

Figure 22. Distribution of surprisal values for pronominal adverbs per 50 years in SG.

In German, all relativizers become less surprising over time and the drops are steeper in scientific discourse than in general language. For the German relativizer (*d-*), contextual predictability is fairly similar between the two subcorpora (GG, figure 17 and SG, figure 20). Median surprisal values drop towards 1750 and stabilize between 1750 and 1900 at around four bits, which is the lowest surprisal value amongst all three relativizer types observed in this study. This is due to the fact that (*d-*) is the overall most frequent relativizer in both GG and SG. We can observe that the whiskers of surprisal values for (d-) in 1650 are rather broad and become narrower over time. This suggests a conventionalization of contexts of the relativizer. Looking at *(welch-)*, (figures 18 and 21), we see that the ranges of surprisal values first expand towards 1750, indicating that *(welch-)* first becomes increasingly variable in terms of contexts. This development initially seems to unfold simultaneously to its increase in frequency. After 1750, surprisal ranges become narrower, indicating a conventionalization of the contexts. In GG, the median surprisal, however, stays fairly stable over time while in SG surprisal steadily goes down. The overall frequency increase of *(welch-)* in SG obviously brings with it an increase in total contexts the relativizer occurs in. The gradual decrease in average surprisal, however, reflects that the contexts become more similar leading to easier predictability of the target word. The contextual predictability of pronominal adverbs shows a similar decrease over time, dropping lower in SG (figure 22) than in GG (figure 19), indicating conventionalization of contexts in SG. Comparing trends in German and English, we find that English relativizers overall become more surprising while in German they become less so. This is primarily due to the overall development of RCs becoming less frequent in English than in German. Only *which* in SE is stable in surprisal and seems to occur in highly predictable contexts. In the next section, we perform qualitative analyses of the syntagmatic environments of relativizers to gain further insights on the contexts RCs tend to occur in.

5.4 Syntagmatic context of relativizers

In the following qualitative analysis, we concentrate on the relativizer *which* in English and *(welch-)* in German, since these have shown to be the most distinctive ones for scientific discourse.

5.4.1 *Grammatical contexts*

To find out what contexts the relativizers occur in and whether these change over time, we first extract all part-of-speech trigrams preceding *which*[2] and *(welch-)*[3] and plot the three most frequent trigrams in each time period. Since the most frequent three in one period may overlap with trigrams from other periods, the total number of trigrams displayed varies. A low total number of trigrams in each figure indicates a lower variation between periods, while a higher number points to stronger variation. Our hypothesis here is that contexts in scientific language become more conventionalized, i.e., the frequencies of a specific pattern surpass the frequencies of other patterns.

[2] Penn Tagset: DT = determiner, NN = noun, IN = preposition, JJ = adjective

[3] STTS: ART = article, NN = noun, APPR = preposition, PT = punctuation (In earlier stages, German punctuation was not standardized yet, thus for this study all punctuation marks (<,>, <;>, </>, etc.) were normalized to 'PT' for better comparability.)

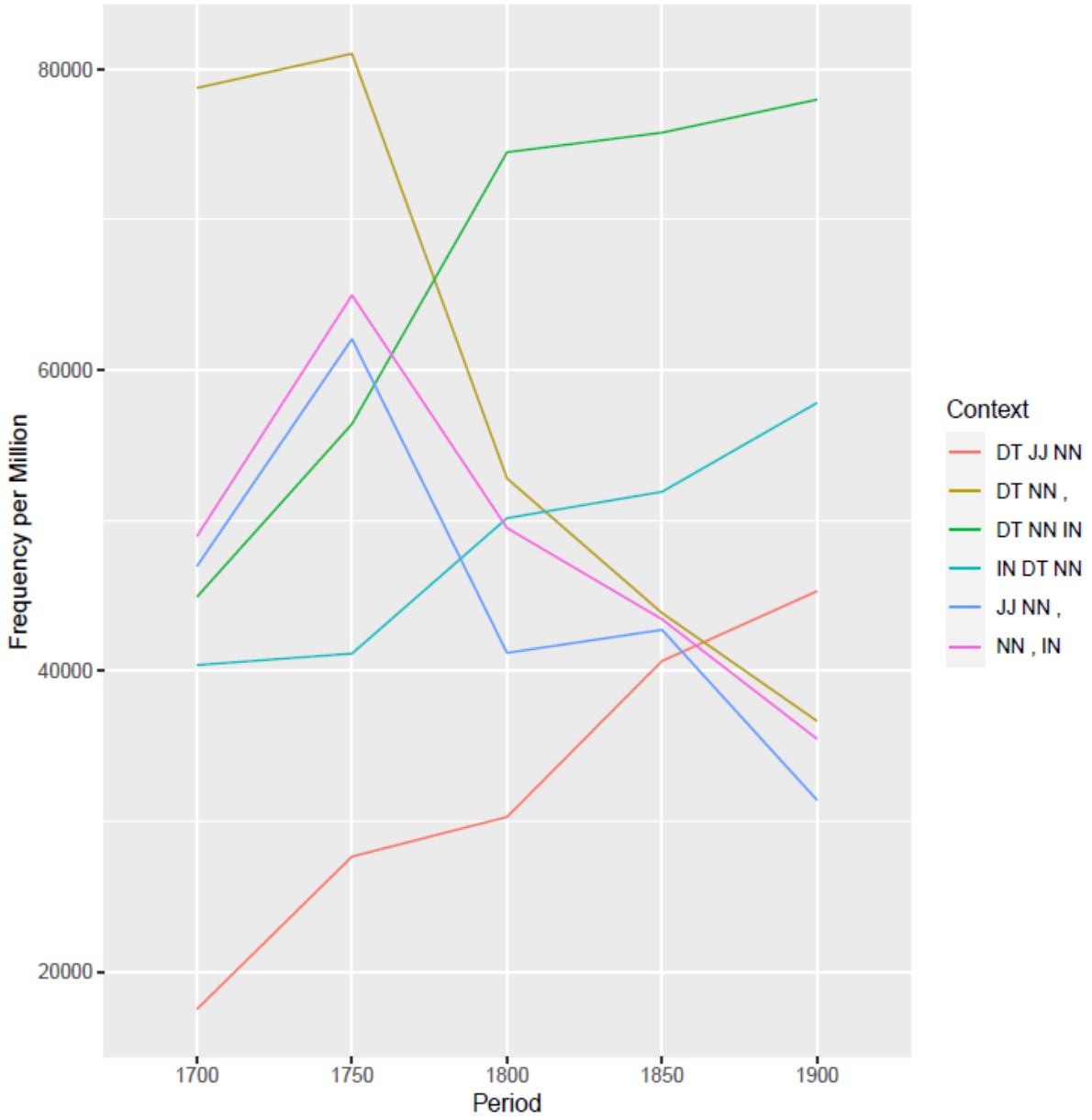

Figure 23. Three most frequent part-of-speech trigrams preceding “which” in GE per 50 years.

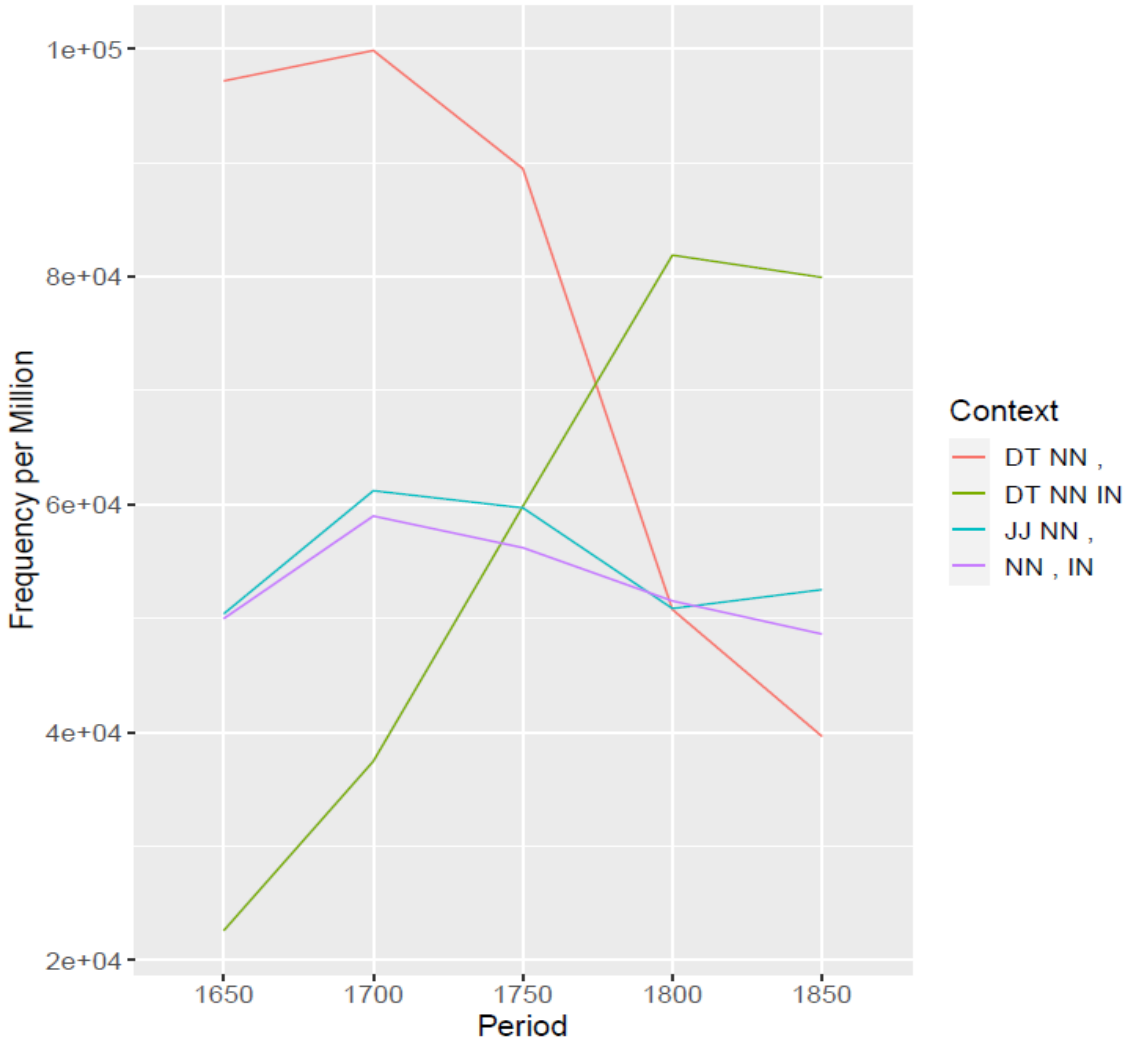

Figure 24. Three most frequent part-of-speech trigrams preceding "which" in SE per 50 years.

Comparing the trigram contexts in GE (figure 23) vs. SE contexts (figure 24), we can report a higher variation between most frequent contexts preceding relativizers in GE, as well as a more diverse set of increasing frequent contexts in GE than in SE. In the scientific corpus, all contexts decrease except for one clearly preferred pattern <*determiner noun preposition*> (DT NN IN), representing RCs introduced by a stranded preposition. Altogether, this corroborates our assumption that, in scientific discourse, contexts of RCs become increasingly conventionalized, again in line with Degaetano-Ortlieb and Teich (2019) and Teich *et al.* (2021). In German (figure 25 for GG; figure 26 for SG), the contexts preceding *(welch-)* are less diverse than in English, showing three clearly preferred contexts in both subcorpora, <adjective noun punctuation mark> (ADJA NN PT) representing shorter nominal phrases, <article noun punctuation mark> (ART NN PT) representing longer nominal phrases and <noun punctuation mark preposition> (NN PT APPR). The fact that these three patterns are continuously amongst the three most frequent POS contexts in both subcorpora alike points to a relatively rigid grammatical environment of RCs in German compared to English. Also, in both GG and SG the trajectories of the trigrams' normalized frequencies are strikingly similar until 1800, all of them increasing.

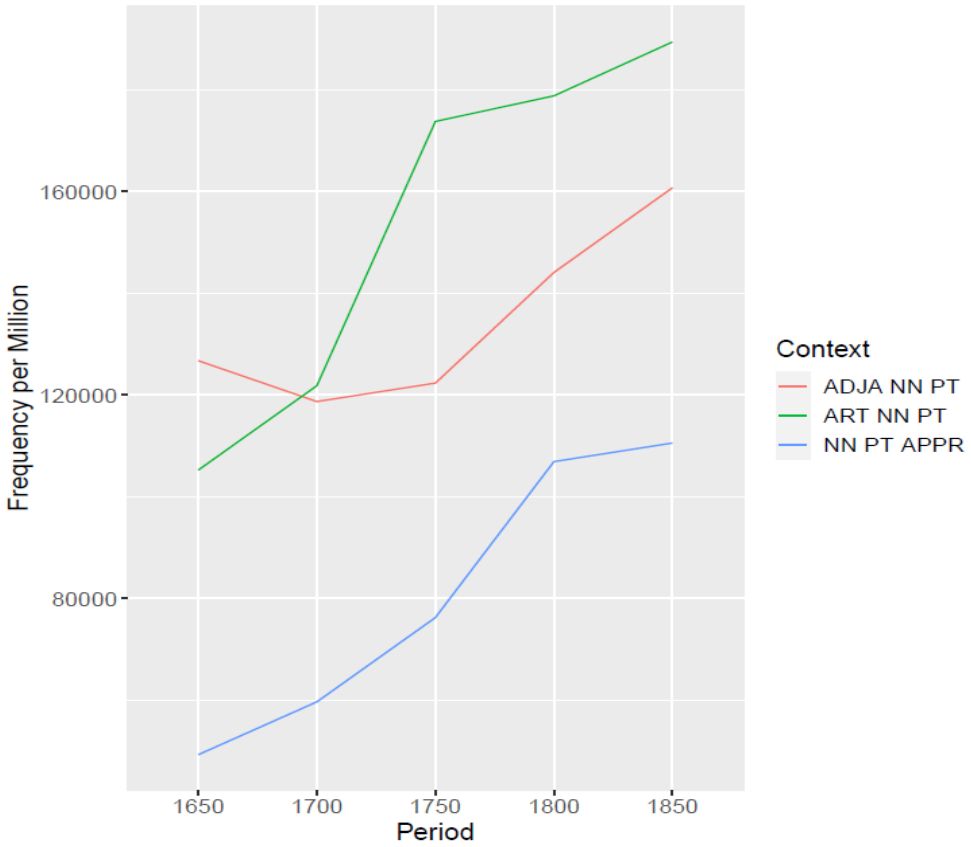

Figure 25. Three most frequent part-of-speech trigrams preceding (*welch*-) in GG per 50 years.

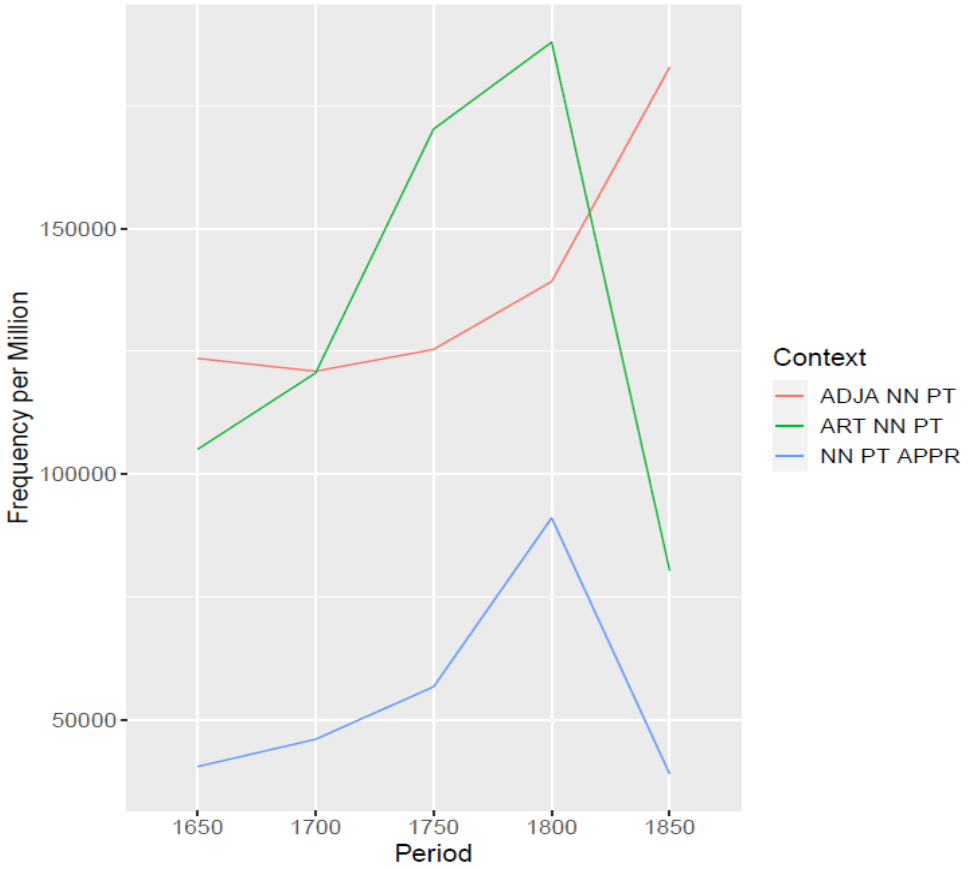

Figure 26. Three most frequent part-of-speech trigrams preceding (welch-) in SG per 50 years.

Between 1850 and 1900, however, in SG only the complex noun phrase pattern <adjective noun punctuation> (ADJA NN PT) increases, while the other two decrease. This again points to a strong differentiation in contextual use of *(welch-)*, similar to the observed trend of *which* in SE.

5.4.2 *Lexical contexts*

We further analyze the most frequent lexical trigrams preceding relativizers in both languages and subcorpora. Tables 2−5 show the three most frequent lexical trigrams per period and specific types they can be grouped in. The types occurring most often in a subcorpus are in boldface.

In English, the top three lexical trigrams preceding *which* show two clear tendencies. In GE (table 2), most top three trigrams (9/15) describe manner expressions (*the manner in*; *the way in*), forming complex conjunctions. In SE (table 4), the trigrams appearing most often are expressions of quantification (8/15) (*, out of*; *, some of*; *, one of*). In 1750 and 1850, however, the top most frequent trigram preceding *which* is the complex conjunction expressing manner, (*the manner in*), matching the top most frequent POS trigram (DT NN APPR). Interestingly, both general and scientific texts show a preference for manner expressions followed by *which*, while the exact lexical form differs between subcorpora. In 1850 GE, (*the way in*) and (*the manner in*) still compete, giving way to the first option from 1900 onwards. SE, however, adopts (*the manner in*).

Table 2. Lexical 3gram context of "which" in CLMET.

period	Freq pM	Freq raw	3-gram	type
1700	2648.87	52	, and of	partitive
	1782.89	35	the manner in	**manner**
	1579.14	31	, and to	-
1750	827.31	106	the manner in	**manner**
	556.85	82	in consequence of	causal
	493.21	78	, one of	quantification
1800	836.32	178	the manner in	**manner**
	562.91	81	, and in	-
	498.58	79	the way in	**manner**
1850	1074.60	137	the way in	**manner**
	723.29	99	the manner in	**manner**
	640.63	63	, all of	quantification
1900	3254.27	54	the way in	**manner**
	2190.37	32	the sense in	**manner**
	1940.05	19	in a way	**manner**

Table 3. Lexical 3gram context of "which" in RSC.

period	Freq pM	Freq raw	3-gram	type
1650	2882.78	62	, out of	**quantification**
	1999.35	43	the doing of	-
	1673.87	36	of it ,	-
1700	2686.05	75	, some of	**quantification**
	2471.17	69	of it ,	-
	2184.66	61	, one of	**quantification**
1750	2359.34	118	the manner in	manner
	1739.51	87	, one of	**quantification**
	1719.52	86	, some of	**quantification**
1800	4168.46	320	the manner in	manner
	1875.81	144	, one of	**quantification**
	1706.46	131	the mode in	manner
1850	2884.18	228	the manner in	manner
	1922.79	152	, each of	**quantification**
	1846.89	146	, one of	**quantification**

In German, the majority of the top three lexical trigrams preceding *(welch-)* for both registers (tables 4−5) are pronominal antecedents, i.e. (*von denen,*). The lexical trigrams do not match the most frequent POS trigrams representing noun phrases, which suggests that, in German, RCs do not tend to occur in conventionalized contexts. Apart from pronominal antecedents, we find prepositional clauses (*Zeit*, *in*; as in example (4a) and (4b) below), as well as semantically empty lexical bundles of grammatical words (*ist es,*) or (*zu sein,*). Taking a closer look at the full contexts of those verbal fragments, i.e. (*ist es,*) in the 19th century German RCs, we find the antecedents to be topicalized noun phrases in cleft constructions, as illustrated in example (5).

Table 4. Lexical 3gram context of (*welch*-) in GE.

period	Freq pM	Freq raw	3-gram	type
1650	2453.81	51	die jenige /	pronoun
	1395.30	29	Wurzel / aus	preposition
	1347.19	28	den jenigen /	pronoun
1700	4684.03	173	, Fluß,	NP
	1868.20	69	von denen,	pronoun
	1272.54	47	in England,	PP
1750	2807.97	62	zu machen,	*to*-infinitive
	1856.88	41	. Diejenigen,	pronoun
	1585.14	35	als die,	pronoun
1800	1625.54	33	als die,	pronoun
	1428.50	29	Zeit, in	preposition
	1083.69	22	ist, in	preposition
1850	1661.13	42	Zeit, in	preposition
	1067.87	27	ist es,	cleft
	909.67	23	Tage, an	preposition

Table 5. Lexical 3gram context of (*welch*-) in SE.

period	Freq pM	Freq raw	3-gram	type
1650	3528.29	52	die jenige /	pronoun
	1899.85	28	die jenigen /	pronoun
	1832.00	27	der Linie /	NP
1700	1413.27	38	Tochter, mit	preposition
	1264.50	34	daß diejenigen,	pronoun
	1115.74	30	zu sehen,	*to*-infinitive
1750	2644.88	157	. Diejenigen,	pronoun
	1212.94	72	zu sein,	*to*-infinitive
	1179.25	70	als die,	pronoun
1800	2967.13	157	. Diejenigen,	pronoun
	1360.72	72	zu sein,	*to*-infinitive
	1322.93	70	als die,	pronoun
1850	1567.25	194	ist es,	cleft
	1147.16	142	Zeit, in	preposition
	1009.82	125	sind es,	cleft

The trigram (*zu sein,*), illustrated in example (6), derives from *scheinen* + *zu*-infinitive constructions (Engl. *seem* + *to*-infinitive).

(4)

a) Nach dem vorigen ist die *Zeit, in welcher* das ganze Gefäß ausfließt, T = [formula]. (gerstner_mechanik02_1832, Scientific German)

b) Endlich nahte die *Zeit, in welcher* man in den Sternenhof gehen sollte. (stifter_nachsommer02_1857, General German)

(5) Gerade diese Zugehörigkeit zu einer und derselben Gruppe von Vorstellungen ist es, *welche* hier die Annahme von Ähnlichkeitsassoziationen rechtfertigt. (kraepelin_arzneimittel_1892)

(6) Es scheint mir ferner eine berechtigte Auffassung zu sein, *welche* Darwin in einem trefflichen Beispiele ausspricht […] (roux_kampf_1881)

Overall, *(welch-)* shares similar lexical contexts in both subcorpora with a slight preference for cleft- constructions in SG and a slight preference for prepositional phrases in GG. The fact that lexical trigrams do not map the POS trigrams shows that, in German, RCs are not introduced by lexicalized multi-word units. Instead, they often occur in frequent syntactic constructions such as *to*-infinitives and topicalization in cleft-constructions. The decreasing surprisal values of German relativizers (which are calculated on lexical patterns) are most likely attributable to the increasingly conventionalized use of comma introducing relative clauses.

6. Summary and discussion

In this paper, we have conducted a comparative study on German and English relativizers as indicators of grammatical complexity. Specifically, we pursued the hypothesis that scientific texts become grammatically less complex compared to texts from general language. We tested for complexity in terms of syntactic intricacy, paradigmatic richness and contextual predictability of relativizers.

Our hypothesis that RCs become less frequent in scientific language is confirmed. However, this development is not exclusive to scientific language but rather concerns all subcorpora. While the decrease in English follows a linear trend, in German RC frequency first increases immensely until the second half of the 18th century and only decreases afterwards, perfectly in line with descriptions by Möslein (1974) and Beneš (1981).

In terms of embeddedness, we found that in English indeed, the average number of RC embeddings per sentence decreases proportionally to the overall number of relativizers in a 50 years' period. The trend confirms Halliday and Martin's (1993) claims about a reduction in syntactic intricacy in scientific English. In German, we found that embeddedness is overall stronger in general German than in scientific discourse. Embeddedness in German is highest before RC frequencies reach their climax, indicating a trend towards a more balanced subordination over time. Overall, a comparison of mere frequencies did not show a register specific trend for lower syntactic intricacy in terms of RC use in the observed time periods, but rather a general linguistic development during the time between 1650 and 1850.

In terms of paradigmatic richness, we found that scientific English developed from a richly populated paradigm of manifold relativizers (especially pronominal adverbs) in 1650 towards a clear preference for one single relativizer, *which*, in 1850. General English, in contrast, remained stable, showing broadly the same distributions of relativizers across all time periods. Calculating entropy, we found that scientific English shifted towards an extremely low uncertainty about an upcoming relativizer, which confirms Degaetano-Ortlieb and Teich's (2019) theory of conventionalization. For German, we found an inverse development. General German develops towards an increasingly confined choice, prioritizing (*d-*) and continuously decreasing in entropy, while scientific German shows a much broader choice of relativizers and consistently high entropy until 1850. In the second half of the 19th century, we observe a drop in use of pronominal adverbs ultimately leading to a decrease in entropy between 1850 and 1900. The findings also show that scientific German over a long stretch of time prefers a rich paradigmatic choice for sophisticated expression, while in English scientific texts a smaller set of choices is preferred. Again, between 1850 and 1900 trends in the two scientific subcorpora align.

Regarding contextual predictability, for English we found overall increasing surprisal for all relativizers in both subcorpora, while in German surprisal values go down. This general result reflects the development of relativizer frequency in the two languages. Obviously, when relativizers overall become less frequent, like in English, they become less predictable. At the same time, *which* becomes least surprising in scientific texts, confirming that during register formation certain words become conventionalized in specific contexts. In German, we see an inverse development. All relativizers become more predictable due to their increasing frequency over time. However, we observed steeper drops in surprisal in scientific texts, especially for the relativizer *(welch-)*. Surprisal values of *(welch-)* also show a smaller range indicating increasingly conventionalized contexts of the relativizer most strongly associated with scientific discourse.

Our qualitative comparison of grammatical and lexical contexts of the relativizers *which* and *(welch-)* showed that in English the most frequent grammatical and lexical contexts of *which* overlap and represent highly lexicalized multi-word units (i.e. (DT NN IN) expressing manner and quantification (*the manner in which, one of which*). In German, the most frequent grammatical contexts do not match with most frequent lexical contexts, indicating that grammatical contexts in German do not become lexicalized over time. The most frequent lexical contexts rather reflect common grammatical constructions of the time, such as topicalized cleft-constructions (*Die Frau war es, welche den Mann schlug.*) and *to*-infinitives in epistemic phrases (*scheint eine Frau zu sein, welche…*) typical for scientific discourse.

Overall, in the scientific subcorpora, we found largely inverse developments in English (becoming less complex) and German (becoming more complex) until the first half of the 19th century and an alignment towards lower complexity in the second half. The results are in line with related work (Möslein, 1974; Beneš, 1981; Admoni, 1990) and our hypothesis that grammatical complexity in German should decrease much later than in English. The delayed shift towards lower complexity in German scientific language may be due to several factors,

such as the longstanding Latin influence on German linguistic style (cf. Habermann, 2001), as well as a much later institutional implementation of German scientific discourse and ultimately a language specific preference for explicit style as compared to English (cf. House, 2006).

Acknowledgements

I am grateful to the anonymous reviewers for their detailed comments and suggestions for improving the paper.

References

Aarts, B., López-Couso, M.J. and Méndez-Naya, B. 2012. Late modern English syntax. In *Historical Linguistics of English*, A. Bergs and L.J. Brinton (eds), 869–887. Berlin/Boston: Mouton de Gruyter.

Admoni, W. 1990. *Historische Syntax des Deutschen*. Tübingen: Niemeyer.

Ágel, V. 2000. Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts. In *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*, W. Besch, A. Betten, O. Reichmann and S. Sonderegger (eds), 1855–1903. Berlin: De Gruyter.

Ball, C. 1996. A Diachronic Study of Relative Markers in Spoken and Written English. *Language Variation and Change* 8 (2): 227−258.

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

Betten, A. 2016. *Grundzüge der Prosasyntax*. Berlin/Boston: Max Niemeyer Verlag.

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

Biber, D. 1993. The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities* 26 (5-6): 331−345.

Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers.* Amsterdam: John Benjamins Publishing.

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

Biber, D. and Clark, V. 2002. Historical Shifts in Modification Patterns with Complex Noun Phrase Structures. In *English Historical Syntax and Morphology. Selected Papers from 11 ICEHL, Santiago de Compostela 2002*, T. Fanego, J. Pérez-Guerra and M.J. López-Couso (eds). 43−66.

Biber, D. and Conrad, S. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D. and Finegan, E. 1997. Diachronic Relations among Speech-based and Written Registers in English. In *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, T. Nevalainen and L. Kahlas-Tarkka (eds), 253−275. Helsinki: Modern Language Society.

Biber, D. and Gray, B. 2011. Grammatical Change in the Noun Phrase: The Influence of Written Language Use. *English Language and Linguistics* 15 (2): 223−250.

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

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

Brooks, T. 2006. *Untersuchungen zur Syntax in oberdeutschen Drucken des 16.–18. Jahrhunderts*. Frankfurt a.M.: Lang.

Crocker, M.W., Demberg, V. and Teich, E. 2015. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz* 30 (1): 77−81.

Dal, I. 2014. *Kurze deutsche Syntax auf historischer Grundlage*. Berlin: De Gruyter.

Degaetano-Ortlieb, S., Kermes, H, Khamis, A. and Teich, E. 2016. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In *Selected Papers from Varieng − From*

Data to Evidence, *Language and Computers*, C. Suhr, T. Nevalainen and I. Taavitsainen (eds), 258−281. Leiden: Brill.
Degaetano-Ortlieb, S. and Teich, E. 2016. Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 165−173.
Degaetano-Ortlieb, S. and Teich, E. 2019. Toward an Optimal Code for Communication: The Case of Scientific English. *Corpus Linguistics and Linguistic Theory*. Available at https://www.degruyter.com/document/doi/10.1515/cllt-2018-0088/html [Last accessed 2 June 2021].
Diller, H., De Smet, H. and Tyrkkö, J. 2011. A European Database of Descriptors of English Electronic Texts. *The European English Messenger* 19: 21−35.
Ebert, R.P. 1986. *Historische Syntax des Deutschen II: 1300–1750.* Bern: Lang.
Fleischer, J. 2004. A Typology of Relative Clauses in German Dialects. In *Trends in Linguistics. Dialectology Meets Typology. Dialect Grammar from a Cross-linguistic Perspective*, B. Kortmann (ed.), 211–243. Berlin/New York: De Gruyter Mouton.
Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C. and Wiegand, F. 2018. Das Deutsche Textarchiv als Forschungsplattform historische Daten in CLARIN. In *Digitale Infrastrukturen die germanistische Forschung (= Germanistische Sprachwissenschaft um 2020, Bd. 6*), H. Lobin, R. Schneider and A. Witt (eds), 219–248. Berlin/Boston: De Gruyter.
Görlach M. 2004. *Text Types and the History of English*. Berlin/New York: Mouton de Gruyter.
Guy, G. and Bayley, R. 1995. On the Choice of Relative Pronouns in English. *American Speech* 70 (2): 148–162.
Habermann, M. 2011. *Deutsche Fachtexte der frühen Neuzeit*. Berlin/Boston: De Gruyter.
Halliday, M.A.K. and R. Hasan. 1985. *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective.* Oxford: Oxford University Press.
Halliday, M.A.K. 1988. On the Language of Physical Science. In *Registers of Written English: Situational Factors and Linguistic Features*, M. Ghadessy (ed.), 162−177. London: Pinter.
Halliday, M.A.K and Martin, J.R. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer Press.
Hinrichs, L., Szmrecsanyi, B. and Bohmann, A. 2015. Which-hunting and the Standard English RC. *Language* 91 (4): 806–836.
House, J. 2006. Communicative Styles in English and German. *European Journal of English Studies* 10 (3): 249−267.
Hundt, M., Denison, D. and Schneider, G. 2012. Relative Complexity in Scientific Discourse. *English Language and Linguistics* 16 (2): 209–240.
Juzek, T.S., Krielke, M.-P. and Teich, E. 2020. Exploring Diachronic Syntactic Shifts with Dependency Length: The Case of Scientific English. In *Proceedings of the fourth workshop on Universal Dependencies*, Barcelona, Spain.
Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and Teich, E. 2016. The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia.
Krielke, M.-P., Fischer, S., Degaetano-Ortlieb, S. and Teich, E. 2019. System and Use of *wh*-relativizers in 200 years of English Scientific Writing. In *Proceedings of the 10th International Corpus Linguistics Conference 2019.* Cardiff, Wales, UK.
Leech, G., Hundt, M., Mair, C. and Smith, N. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.
Lehmann, H. 2001. Zero Subject Relative Constructions in American and British English. *Language and Computers* 36 (1): 163–177.
Levey, S. 2006. Visiting London Relatives. *English World-Wide* 27 (1): 45–70.
Levy, R. 2008. Expectation-based Syntactic Comprehension. *Cognition* 106 (3): 1126−1177.
Mair, C. 2006. *Twentieth-century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.
Mellinkoff, D. 2004. *The Language of the Law*. Eugene: Resource Publications.
Milin, P., Kuperman, V., Kostic, A. and Baayen, H.. 2009. Paradigms Bit by Bit: An Information Theoretic Approach to the Processing of Paradigmatic Structure in Inflection and Derivation. In

Analogy in Grammar: Form and Acquisition, J.P. Blevins and J. Blevins (eds), 214−252. Oxford: Oxford University Press.

Möslein, K. 1974. Einige Entwicklungstendenzen in der Syntax der wissenschaftlich technischen Literatur seit dem Ende des 18. Jahrhunderts. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 94: 156−198.

Nevalainen, T. 2012. Reconstructing Syntactic Continuity and Change in Early Modern English Regional Dialects: The Case of *who*. In *Analyzing Older English*, D. Denison, R. Otero, C. McCully and E. Moore (eds), 159–184. Cambridge: Cambridge University Press.

Nevalainen, T. and Raumolin-Brunberg, H. 2002. The Rise of Relative *who* in Early Modern English. In *Relativisation on the North Sea Littoral*, P. Poussa (ed.), 109–121. Munich: Lincom Europa.

Nevalainen, T. and Raumolin-Brunberg, H. 2012. Its Strength and the Beauty of it: The Standardization of the Third Person Neuter Possessive in Early Modern English. In *Towards a Standard English*, D. Stein and I. Tieken-Boon van Ostade (eds), 171−216. Berlin/Boston: De Gruyter Mouton.

Österman, A. 1997. *There*-compounds in the History of English. *Topics in English Linguistics* 24: 191−276.

Pickl, S. 2020. Factors of Selection, Standard Universals, and the Standardisation of German Relativisers. *Lang Policy 19:* 235−258.

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

Reichmann, O. and Wegera, K.-P. 1993. *Frühneuhochdeutsche Grammatik*. Tübingen: Niemeyer.

Romaine, S. 1980. The RC Marker in Scots English: Diffusion, Complexity, and Style as Dimensions of Syntactic Change. *Language in Society* 9 (2): 221–247.

Romaine, S. 1982. *Sociolinguistic Variation in Speech Communities*. London: Edward Arnold.

Rubino, R., Degaetano-Ortlieb, S., Teich, E., and van Genabith, J. 2016. Modeling Diachronic Change in Scientific Writing with Information Density. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*: 750−761.

Santorini, B. 1990. *Part-of-speech Tagging Guidelines for the Penn Treebank Project* (3rd revision). Technical Report MS-CIS-90-47, University of Pennsylvania, Department of Computer and Information Science.

Shannon, C.E. 1949. *The Mathematical Theory of Communication*. Urbana/Chicago: University of Illinois Press, 1983 edition.

Tagliamonte, S. 2002. Variation and Change in the British Relative Marker. In *Relativisation on the North Sea Littoral*, P. Poussa (ed.), 147–165. Munich: Lincom Europa.

Tagliamonte, S., Smith, J. and Lawrence, H. 2005. No Taming the Vernacular! Insights from the Relatives in Northern Britain. *Language Variation and Change* 17 (1): 75–112.

Teich, E., Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H. and Lapshinova-Koltunski, E. 2016. The Linguistic construal of Disciplinarity: A Data Mining Approach Using Register Features. *Journal of the Association for Information Science and Technology (JASIST)* 67 (7): 1668−1678.

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S. and Bizzoni, Y. 2021. Less is More/More Diverse: On the Communicative Utility of Linguistic Conventionalization. *Frontiers in Communication* 5. https://doi.org/10.3389/fcomm.2020.620275 [Last accessed 2 June 2021].

Thielen, C., Schiller, A., Teufel, S. and Stöckert, C. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [Last accessed 2 June 2021].

Tottie, G. and Harvie, D. 2000. It's All Relative: Relativization Strategies in Early African American Vernacular English. In *The English History of African American English*, S. Poplack (ed.), 198–230. Oxford: Blackwell.

Voigtmann, S. and Speyer, A.. 2020. *Information Density as a Factor for Syntactic Variation in Early New High German*, LE, Tübingen.

Von Polenz, P. 1999. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart* (Vol. 3). Berlin/New York: Walter de Gruyter.

Author's address

Marie-Pauline Krielke
Department of Language Science and Technology
Saarland University
Building A2.2, room 1.02
Campus
DE-66123 Saarbrücken
Germany
mariepauline.krielke@uni-saarland.de

On brackets in translation (or how to elaborate in brackets)

Magnus Levin, Jenny Ström Herold

Linnaeus University (Sweden)

This paper presents findings on the use of brackets in original texts and translations based on the Linnaeus University English-German-Swedish corpus (LEGS). The results show that in originals, brackets are the most frequent in English and the least in Swedish. Translations usually contain more brackets than originals. There are two reasons for this. First, most brackets are retained, and secondly, many are added. Added brackets mostly contain short synonyms facilitating target-reader comprehension. English translators introduce the most changes (additions, omissions, downgrades and upgrades), and Swedish ones the least. Brackets tend to fulfil content-oriented rather than interpersonal functions. When brackets are replaced by other punctuations marks in translations, these tend to be commas or no punctuation marks at all. German originals have a stronger preference for bracketing phrases than clauses compared to English and Swedish. These German phrasal brackets are often expanded into clauses in translations.

Keywords: brackets, punctuation, LEGS, explicitation, translation strategies, clause building, English/German/Swedish

1. Introduction

Writers of non-fiction are faced with the complex task of conveying complex states of affairs, while simultaneously avoiding making their texts too long. Translators, in turn, often feel the need to make their texts slightly more elaborate to target-text readers by, for instance, adding information on cultural features that are less known in the target culture. Brackets enable the insertion of more or less information-dense additions and would therefore seem to be suitable structures to use for such elaborations. Illustrative examples of bracket usage in an English original and its German translation are given in (1), involving the retention, addition and omission of information:

(1) During the Miocene period *(23–5.3 million years BP)*, the equines diversified and took on the appearance of modern species. The modern survivors of *the equines*, which include horses, donkeys, asses, zebras, kiangs and onagers, evolved during the Pleistocene period *(2.5 million–12,000 years BP)* alongside our own human ancestors *(see the next entry for a more detailed discussion of the early equids)*. [LEGS; English original]

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.v11i1.3441.

> Im Lauf des Miozäns *(vor 23–5,3 Mio. Jahren)* diversifizierten sich die Pferde und ähnelten im Aussehen bereits den heutigen Arten. Die modernen Vertreter *der Pferde (Gattung Equus)*, zu denen Pferde, Esel, Zebras, Tibet-Wildesel und Halbesel gehören, entwickelten sich während des Pleistozäns *(vor 2,5 Mio.–12 000 Jahren)*, gleichzeitig mit den Vorfahren des modernen Menschen Ø. [German translation]

Both the original and the German translation contain three pairs of brackets, but these only partly match each other. The years for the two time periods are transferred directly, while the German translator once adds a term in brackets where the original makes do with only one (i.e., *equines* > *Pferde (Gattung Equus)* ['horses (genus equus)']). Finally, the English signpost (*see the next entry …*) is omitted in the German version, suggesting that German writing sometimes may prefer a less reader-oriented style than English, a hypothesis that will be explored further in this paper.

In the following, we explore both distributions and uses of (round) brackets in English, German and Swedish original and translated popular non-fiction while addressing the following research questions:

- How frequent are brackets in English, German and Swedish originals and translations?
- What functions do brackets serve and what syntactic forms does bracketed text have in originals and translations?
- How are brackets rendered in translations in terms of being, e.g., retained, added or omitted, and what other punctuation marks are used as correspondences?
- To what extent do translations adhere to the target-language norms and/or to what extent does source-text usage "shine through" in translations?

As for the structure of this paper, Section 2 presents the LEGS corpus material and how we went about in the search process. Section 3 gives an overview of the previous studies on punctuation. Section 4 starts by presenting the findings for the originals and then moves on to the patterns observed in translations.

2. Material and method

This study is based on material from the Linnaeus University English-German-Swedish corpus (LEGS) (Ström Herold and Levin, 2019) which includes recently published (2000s) non-fiction books in English, German and Swedish. It is balanced for all three languages and for each original we always include target texts in the other two languages. The corpus covers both narrative and instructive genres, such as biographies, popular science and self-help books. To avoid author- or translator-specific features, each author and translator is represented only once.

The trilingual structure of the LEGS corpus is illustrated in Figure 1:

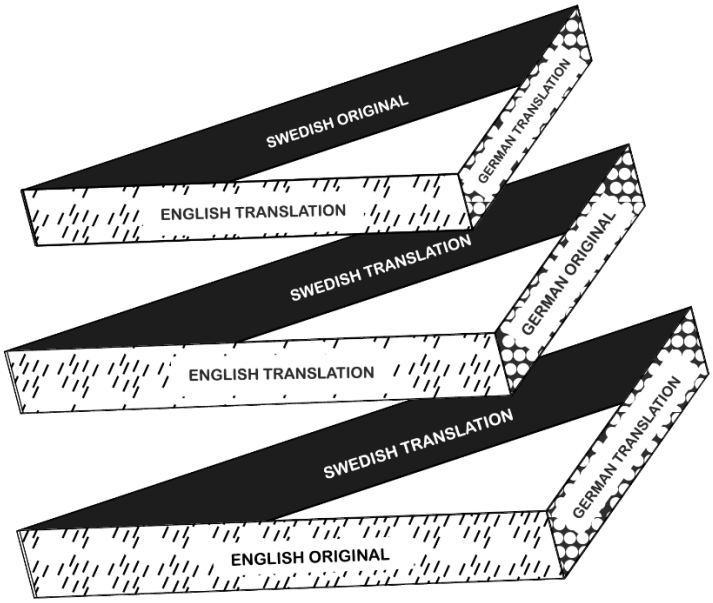

Figure 1. The structure of the Linnaeus University English-German-Swedish corpus (LEGS).

The present study covers a selection of the LEGS corpus, i.e. eight English, eight German and eight Swedish texts with two translations for each text. Some of the texts in the corpus were excluded because they would severely skew the findings. For example, cookbooks were removed due to their extreme use of brackets for measurement conversions both in originals and translations (e.g., *150 g (1 ½ dl) strösocker* (Sw.) > *2/3 cup (5 oz/155 g) sugar*). The bracket usage in this special genre warrants further studies, but the sheer numbers – more tokens in a single text than in a whole subcorpus – would turn this into a study of brackets in cookbooks.[1] The approximate word counts for each subcorpus in the present study are given in Table 1.

Table 1. Word counts for the LEGS subcorpora included in the present study.

		English translation	**German translation**	**Swedish translation**
English originals	434,000	*	416,000	421,000
German originals	329,000	374,000	*	337,000
Swedish originals	335,000	353,000	331,000	*

The table shows some perhaps peculiar differences in word counts. However, these can, at least to some extent, be explained by structural or cultural differences between the languages. For instance, German and Swedish use solid compounds (or compounds with hyphens) while English usually writes noun modifiers separate from their head nouns (e.g., *Rolex watches* > *Rolex-Uhren* (Ge.); *Rolexklockor* (Sw.); see Ström Herold and Levin, 2019), which means that the word count will increase in languages where juxtaposition is prevalent. Culturally motivated additions or omissions, sometimes of whole sections, is an additional factor.

It should be noted that the books available in each source language affect the selection of texts included in the corpus. Not only are the books translated from German and Swedish shorter than those translated from English, more German and Swedish books also tend to belong to more reader-oriented genres, such as instructive self-help books. As will be seen below, the greater proportions of such interpersonal texts in these languages have some bearing on the results.

[1] Another Swedish original text was discarded because the English translator transformed 600 endnotes, mostly containing references, into brackets in the running text.

For our search we used a custom-made LEGS interface and included all round brackets in originals aligned with their corresponding translation segments in the two target languages. We also searched for all round brackets in translations having a "non-bracket" in the original. In all, this procedure retrieved 5923 bracket pairs in either originals or translations as well as 1987 non-bracketed correspondences.[2] We manually checked both originals and translations to ascertain that the extensive use of footnotes or endnotes would not affect the findings. The source-text and target-text instances were classified according to their functional and formal features and, for target texts, the translation strategies that were applied. The approach in the present study thus exploits the two main advantages of using translation corpora rather than monolingual reference corpora when comparing languages as argued by Nádvorníková (2020: 46): first, it allows comparisons between frequencies in originals and translations, which can be considered to be equivalent texts, and second, it enables the analysis of translation strategies and the punctuation systems of different languages.

The following section provides an overview of previous work on brackets and related punctuation marks such as commas and dashes. In general, comparative or translation-based studies on punctuation use in different languages are rare, which also applies to brackets. Nevertheless, some important translation trends have been noted that will be explored further in this study.

3. Brackets in monolingual and contrastive studies

According to Leech *et al*. (2009: 246) brackets (both round and square) "have increased immensely" in English and are typical for a more "serious written style" (cf. also Crystal, 2015: 157). Similarly, Biber and Gray (2016: 120) remark on their frequent use in academic prose, more specifically as information-dense juxtaposed appositions – NP (NP) – as in *International Meta-analysis of mortality Impact of Systemic Sclerosis (IMMISS),* where the acronym is introduced in brackets. In this example, the spelt-out term and the bracketed acronym are co-referential, which, as suggested by Biber and Gray (2016: 205−206), was also how brackets were used originally in English. Nowadays, brackets may encompass all sorts of information. Biber and Gray (2015: 205) show that these may include descriptive specifications or more "distant" information and that the bracketed text can be relatively lengthy and complex, yet nominally dense, as in their example: *Numerous variables were measured, including* […] *date of enrollment (date of first visit to the cohort with the pertinent diagnosis), age at first visit* […]. In a similar vein, Bredel (2018: 11) refers to brackets as "communicative marks". By using brackets, authors make themselves visible in their text (cf. also Baumgarten *et al*., 2008: 188), by illustrating or explaining previous information to the reader: *Sie saßen (es war Winter geworden) in der Stube* ['They sat (it had turned winter) in the living room] (Bredel, 2018: 12). Here, the brackets include a complete sentence which, parenthetically, supplies the background information that the author deemed necessary for the interpretation. It should be noted that brackets are considered optimal candidates for parenthetical inserts (parentheses), appearing medially, but that they may also appear finally in a sentence (Quirk *et al*., 1985: 1625).

In monolingual studies and reference books on punctuation, brackets are sometimes contrasted with other 'correlative' punctuation marks (Quirk *et al*., 1985: 1625−1631), i.e., punctuation pairs. In these sources, either their interchangeability is highlighted or differences

[2] In contrast to Baumgarten *et al*. (2008), we did not include square brackets in our searches, but among our included tokens are some instances of such brackets occurring as correspondences of round brackets. For instance, a German author marking an omission in a quote with round brackets *(...)* was rendered by the English translator in square brackets *[...]*. The few hundred remaining square brackets are not likely to have affected the results decisively.

in terms of style and semantics. For instance, the German *Duden* (RgD, 2016: 517) suggests that parenthetical inserts can be placed between either brackets, commas or dashes, without any stylistic or semantic implications. Bredel (2018: 12−13), on the other hand, suggests that the choice of punctuation marks has semantic bearing. Returning to the example above, replacing the brackets with commas would yield a slightly different interpretation or focus: *Sie saßen, es war Winter geworden, in der Stube*. According to Bredel, the commas would emphasise the "syntactic disintegration" of the parenthesis, which is claimed to be different from the more communicatively used brackets. Comparing brackets and commas, the *English Style Guide* (2016/2019: 12), used by the European Commission, writes that brackets are used much like commas, except that a bracketed text segment, compared to an insert between commas, has "a lower emphasis", i.e., is more strongly backgrounded. From a stylistic perspective, dashes are usually said to have a more informal, dramatic flair than brackets and commas (Quirk *et al*., 1985: 1629; Leech *et al*., 2009: 245; Crystal, 2015: 158).

As noted above, brackets, and punctuation marks in general, have gone much unnoticed in translation studies, which is surprising considering that the appropriate use of punctuation marks is not a trivial matter for translation students or professional translators (cf. Ingo, 2007: 67; Shiyab, 2017: 93−101). However, the last few decades have seen a growing interest among translation scholars with studies on the translational rendition of different punctuation marks (Bystrova-McIntyre, 2007: 137−138; Baumgarten *et al.*, 2008; Englund Dimitrova, 2014; Wollin, 2018; Frankenberg-Garcia, 2019; Ström Herold and Levin, forthcoming). Still, many of these studies have rather strong limitations, often based on small data sets (if any) with some rare exceptions. Based on these previous studies, it is possible to tease apart three typical translation tendencies for punctuation marks. One is direct transfer, as the most common translation strategy, often reaching about 90% (Gustafsson, 2013; Wollin, 2018; Frankenberg-Garcia, 2019; Ström Herold and Levin, forthcoming). The second one is normalization/standardisation, where an exaggerated use of punctuation marks used as a stylistic device tends to be toned down in the translation (Englund Dimitrova, 2014: 96). Finally, there is a tendency for explicitation, i.e. the inclination to spell things out rather than leave them implicit (Baker, 1996: 180). Explicitation is often cited as a Translation Universal, i.e., a feature occurring in translations rather than originals, and that is not the result of source-text interference (Baker, 1993: 243). In the context of punctuation marks, explicitation may be reflected in the replacement of a punctuation mark by lexical material (e.g., a colon being replaced by a connector, as argued by Eskesen and Fuglsang (1998)). Explicitation has also been approached from the perspective of brackets. Baumgarten *et al*. (2008: 190) even suggest that brackets are "typical sites of translational explicitation" as they are frequently used by translators to add information that is not present in the originals.

To our knowledge, Baumgarten *et al*.'s study (2008) is the only corpus study of the use of round and square brackets in originals and translations. Their material includes English originals and their German translations, but also a comparable corpus with German originals, facilitating comparisons with 'translated German'. The texts stem from two different popular science magazines and, thus, their material is less varied than the material used in the present study, which includes various texts in the broader non-fiction genre. Their initial assumption was that German translations would contain more/added brackets as a result of translational explicitation, but also because their German control corpus contains more brackets than the English originals. Their results show that the German translations indeed contain very many added brackets (about 60%), but also, surprisingly, that translators remove original brackets to a very high extent (about 70%). Baumgarten *et al*. (2008: 191−192) conclude that most changes in the German translations are due to adaptation, where translators adapt to the textual conventions of the target language, and not translational explicitation. As for the function of the information included in brackets, their study suggests that English originals use brackets

for subjective, writer-based elaborations to a much higher extent than the German originals (ibid. 2008: 200), see example (2):

(2) Subsequent work by Mundt, Bo Reipurth of the European Southern Observatory in Santiago, Chile, and others *(including me)* showed that …

In contrast, German originals and translations tend to use them for reader-oriented, intertextual references and content-based brackets, e.g., for including specialized terminology and biographical or geographical information (ibid. 2008:20), as in (3):

(3) … an der Universität Newcastle upon Tyne *(England)* …

The next section will present the findings on brackets, in English, German and Swedish originals and translations. By combining comparable and translation data, our material allows us to draw conclusions about language norms and translation-related features.

4. Results

4.1 Brackets in originals

This section presents the frequencies of brackets in English, German and Swedish originals, the functions fulfilled by brackets in those texts, the distributions of the functions across originals and finally the syntactic forms of the bracketed text.

First, the quantitative overview in Figure 2 indicates significant frequency differences between the three source-text corpora, using a log likelihood test:

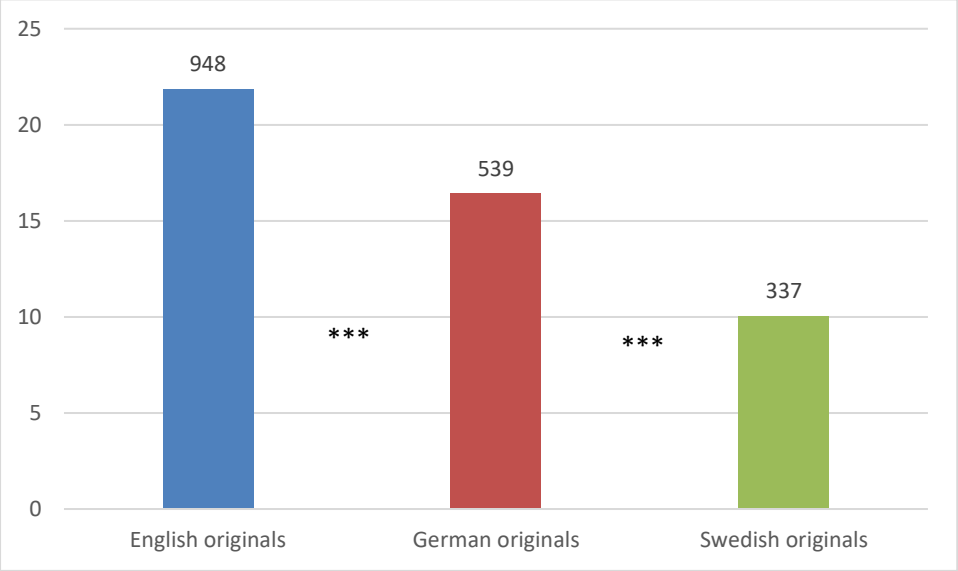

Figure 2. Brackets in English, German and Swedish originals in LEGS per 10,000 words.

The findings seem quite solid with brackets being the most frequent in English originals and the least frequent in Swedish originals. A point illustrating this is that of the seven texts producing more than twenty brackets per 10,000 words, five are English, two German and none Swedish. The two texts with the highest frequencies for German are the two most clearly operative texts (such as a self-help book for cat owners), which suggests that more instructive, reader-oriented genres contain more brackets than more content-oriented genres, such as biographies. The investigated English texts do not include instructive texts, but still yield the highest number of brackets. Further studies of genre-specific uses of brackets are called for to

determine possible differences between more instructive genres and what Leech *et al*. (2009: 245) refer to as more "serious written style".

In the material, we identified two main functional categories, content-oriented brackets and interpersonal brackets, with two subcategories for the former and three for the latter. As the terms imply, content-oriented brackets focus on adding to the content of the text, while interpersonal ones are more reader- or author-oriented and as such may be considered a prototypical case of using brackets as communicative marks (Bredel, 2018: 11). Our labels for the functional categories have been inspired by House's (e.g., 1997, 2011) seminal work on communicative styles in English and German, establishing a cline between the languages with English writing generally being more interpersonal, whereas German is more content-oriented. Interpersonal style is characterized by features such as author presence and reader address, while content-oriented style relies more heavily on transmission of facts.

In the following, the functional subcategories will be presented, starting with the content-oriented synonym and specification, and then the interpersonal reader address, hedge and subjective author comment. Figure 3 presents a graphic overview of the categories and subcategories.

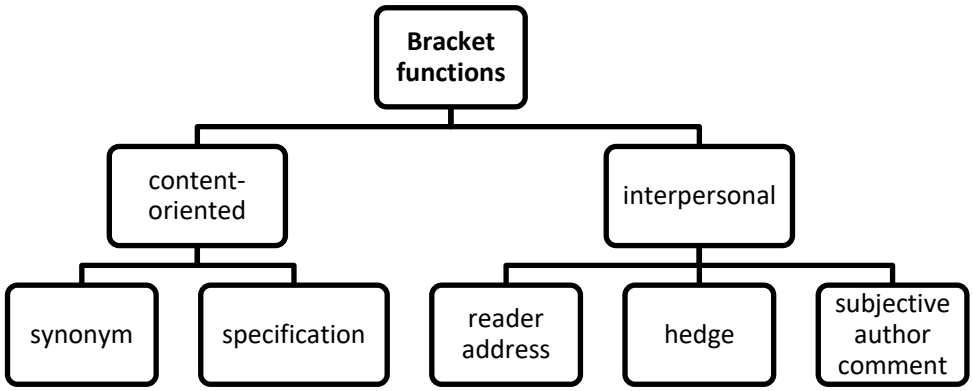

Figure 3. Functional categories of brackets in the LEGS material.

The subcategory synonym relates to the original bracket function proposed by Biber and Gray (2016: 205−206) above. Here, the bracketed text is co-referential with some item outside the brackets. Typical instances involve the addition of name variants (4), measurements using different systems (5), and the introduction of acronyms (6). The second subcategory of content-oriented brackets, i.e., specification, involves the addition of factual details to the non-bracketed text, as exemplified in (7) and (8).

I.I Content-oriented: Synonym

(4) On 3 September at *Bydgoszcz (Bromberg)*, random firing against Poles in the streets led to a massacre … [LEGS; English original]

(5) They undulate their whole body to propel themselves through the water and can reach speeds of *38 km/h (24 mph)*. [LEGS; English original]

(6) The trap was given a name by Dwight Eisenhower in his farewell speech as President on January 17, 1961: *the military-industrial complex (MIC)*. [LEGS; English original]

I.II Content-oriented: Specification

(7) *Imperial Oil (of which Exxon owns a majority share)* sank about $13 billion … [LEGS; English original]

(8) During *the Byzantine period (4th–15th c.)*, the elite preferred expensive silks and linen to woolen garments. [LEGS; English original]

Among the interpersonal brackets, the subcategory of reader address covers instances where the reader is addressed "outside the text", as in (9), and metatextual comments guiding the readers, as in (10). The subcategory of hedges is in some ways categorically ambiguous as hedges may serve slightly different functions. On the one hand, they may refer to the truth value of the content and thus would group with the content-oriented brackets, but on the other hand, they maintain relations with readers, which is why we classified them as interpersonal brackets. This also agrees with Hyland who finds that hedges are often ambiguous and rarely allow just one single interpretation (1996: 437, 439; cited in Kranich, 2011: 82). Thus, example (11) leans more towards a content-oriented interpretation, while (12) is more evidently subjective and reader-oriented in nature. Finally, there are instances where authors subjectively comment on or evaluate facts and events. This is exemplified in (13) and (14) where the subjective stance is highlighted by the adjectives *striking* and *tantalizing*. The examples given also illustrate the different available positions of bracketed texts – most occur sentence-medially, as in (11) – (13), followed by sentence-final position, as in (10), while the rarest is independent sentences, as in (14).[3]

II.I Interpersonal: Reader address

(9) … while the bees are visiting your bee-friendly plants *(if you haven't got any, I hope you'll plant some next spring)* … [LEGS; English original]

(10) The weaknesses of the program have also been hotly debated, particularly the question of whether the decision to phase out nuclear energy has led to a resurgence of coal *(more on that next chapter)*. [LEGS; English original]

II.II Interpersonal: Hedge

(11) Whether the tests on which the participants improved measure perceptual ability, perceptual speed, or *(as the authors interpret it)* stimulus-driven attention is a moot point. [LEGS; English translation from Swedish]

(12) Despite his arrogance *(or perhaps because of it)* he was able to charm Atari's boss. [LEGS; English original]

II.III Interpersonal: Subjective author comment

(13) Andrew *(who bears a striking resemblance to Baldrick from Blackadder)* came up with the cunning plan … [LEGS; English original]

(14) Mists and fogs […] forced Hitler to accept that the Luftwaffe could not provide the vital support needed for his November target date. *(It is tantalizing to speculate how differently things might have turned out if Hitler had launched his attack then rather than six months later.)* [LEGS; English original]

The distributions of functions in originals are given below in Figures 4 and 5. Figure 4 shows the proportions and raw numbers of the two main functional categories, content-oriented and

[3] German has the strongest preference for medial position (74%; English 67%; Swedish 63%), and Swedish for both final position (31%; English 28%; German 23%) and independent sentences (6%; English 5%; German 3%).

interpersonal brackets, and Figure 5 presents the frequencies per 10,000 words for the five subcategories. It should be noted that English originals are presented in the middle of Figure 4 to illustrate the statistical significance.

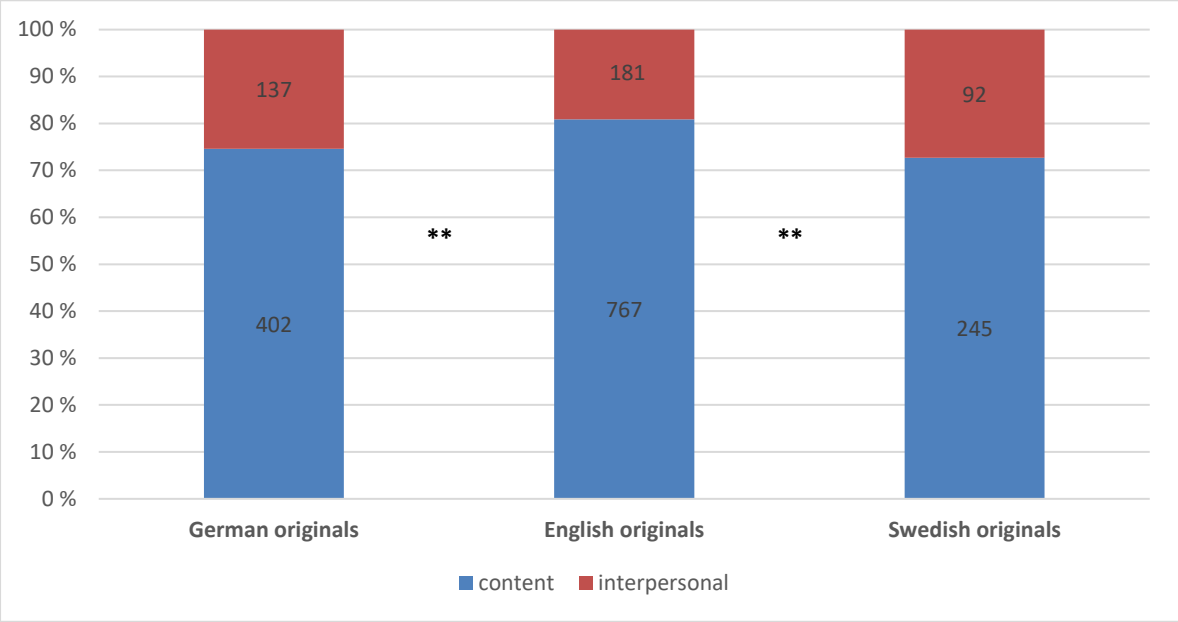

Figure 4. Proportions of main functions: content and interpersonal.

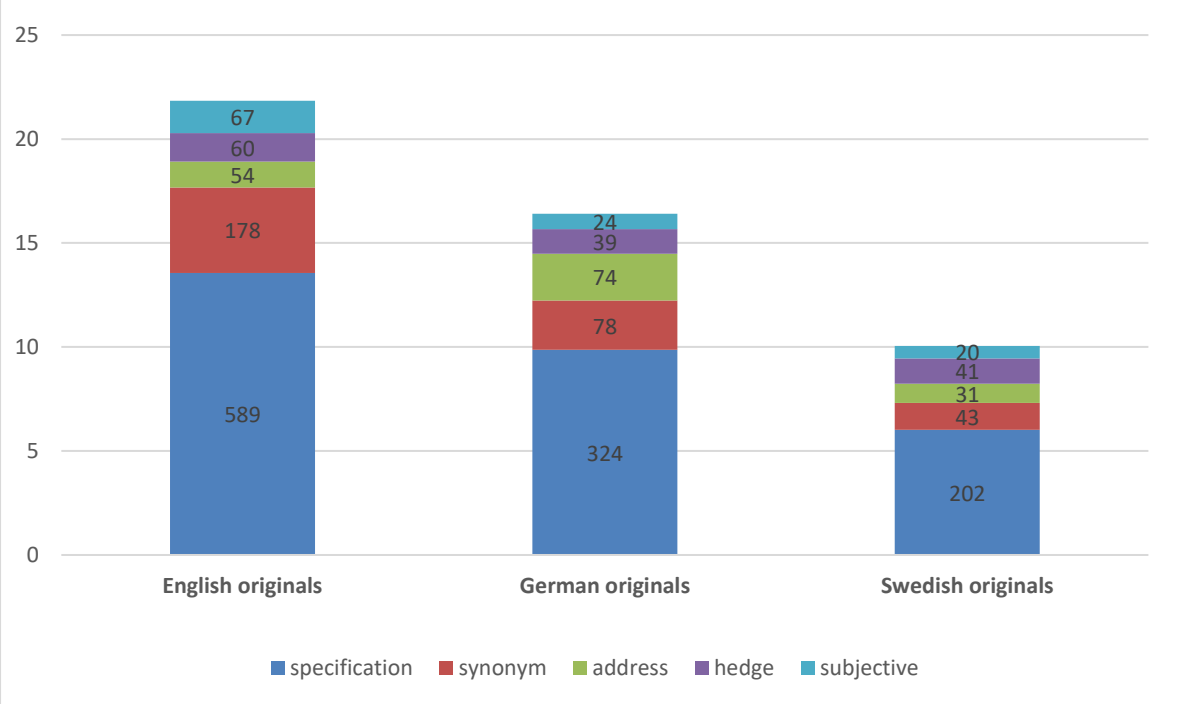

Figure 5. Subcategories of content and interpersonal functions per 10,000 words.

Figure 4 suggests that English original texts are more strongly associated with content-oriented brackets than German and Swedish. This was at least partly unexpected, seeing that Baumgarten *et al.*'s (2018: 191−192) findings on English and German indicate the opposite. In our data, the difference between English and the other languages should nevertheless not be

overemphasized since the effect size is small.[4] Moreover, the lower proportion of reader-oriented genres represented in the English subcorpus in all likelihood promotes a larger proportion of content-oriented brackets in these originals. The functional subcategories in Figure 5 may therefore provide a more relevant picture. Here we see that the two content-oriented subcategories, specification and synonym, are indeed associated with English originals (though not very strongly so).[5] Using signed deviations from expected cell-wise counts and their chi-squared contributions, we notice that English originals show particular dis-preference for the address function.[6] The most significant preference among the functional subcategories is for reader address in German and the second strongest for hedges in Swedish.[7] Synonyms are somewhat more popular in English and somewhat less so in Swedish.[8] The functions specification and subjective author comment are more or less equally preferred in the three languages.

Looking more closely at the observed differences between the subcategories, it is evident that one particular reason for the more frequent use of synonyms in English is the recurring use of synonymous measurements in this language, (as exemplified in (5)), a usage that is absent in the other languages. Moreover, the more frequent use of reader address in German is likely an influence of the German predilection for using intertextual signposts, as also noted by Baumgarten *et al*. (2008: 200) (e.g., *(ab S. 42)* translated into *(see page 42 onward)*). As for hedges, most Swedish instances occur in three books written by professors, which may indicate that bracketed hedges are a particular academic phenomenon carried over into the popular domain.

Apart from the functions of the bracketed texts, we decided to also take a closer look at the forms of the bracketed texts. This focus was inspired by some previously noted differences between the languages, one being the increasing German aversion to subordinate clauses, probably an avoidance strategy for difficult-to-process verb-final clauses (Becher, 2011; Bisiada, 2013; Ström Herold and Levin, 2018, forthcoming) and the other one being the overall German preference for nominal style (cf. Carlsson (2004) for German in contrast with Swedish). Therefore, all brackets were classified according to their syntactic form, either as i) clausal, i.e., instances which contain a verb phrase, or ii) phrasal, i.e., instances which correspond to a noun phrase or a prepositional phrase (examples of clausal brackets are given in, e.g., (7) and (9), and phrasal instances in (4) and (8)). Figure 6 presents the results from our originals:

[4] Independence of English vs Swedish and German collapsed, χ^2=12.58, df=1, p=***; Cramer's V=0.08
[5] Using a cell-by-cell chi-square contributions; 0.32 for specification and 7.21 for synonyms.
[6] Deviation -28.6, cell-wise χ^2 contribution is 9.92
[7] Deviations and cell-wise χ^2 contributions are 27.01 and 15.53, and 15.13 and 8.85.
[8] Deviations and cell-wise χ^2 contributions are 22.59 and 3.29, and -12.24 and 2.71.

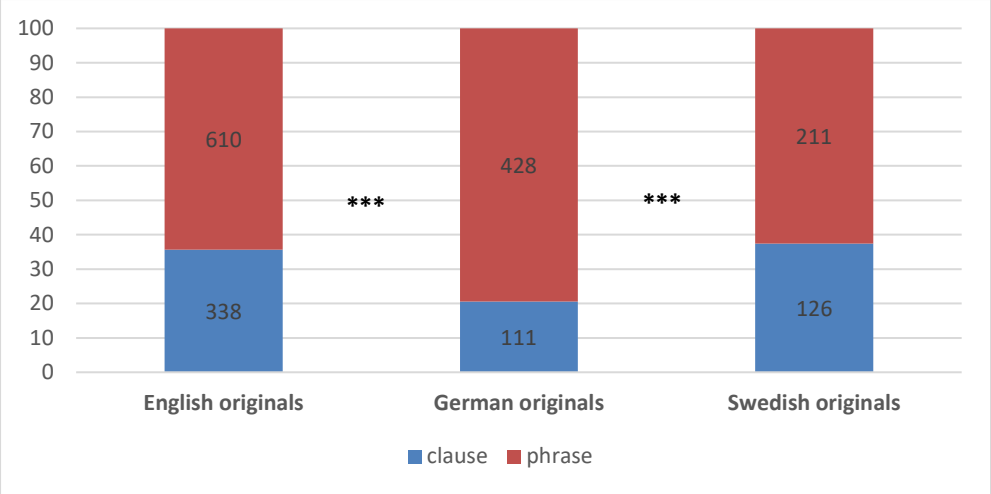

Figure 6. Proportions of clausal and phrasal brackets in English, German and Swedish originals.

The findings show that German originals put clauses in brackets significantly less than the English and Swedish ones. (For the frequencies of clauses and phrases in translations, see Section 4.2.4.). This result is in line with the above-mentioned contrastive studies suggesting that German prefers a more information-dense phrasal style and avoids subordinated clauses. This makes sense, as interpersonal functions such as addressing readers or adding author comments will require more elaborate structures than many content-oriented brackets which consist of, for example, one-word synonyms. Indeed, there is a positive correlation[9] for all three languages between, on the one hand, clauses and interpersonal brackets and, on the other, phrases and content-oriented brackets.

A related observation concerns the number of words in the bracketed text. Phrases tend to be shorter than clauses and this is also reflected in the brackets in the LEGS material. As illustrated in Figure 7 below, English brackets contain more than 50% more words on average than the German, with the Swedish originals in between.

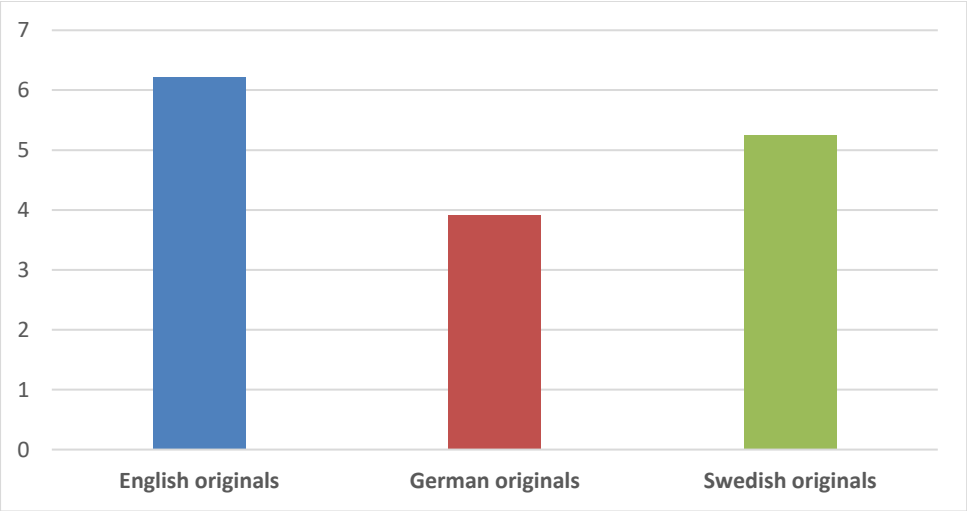

Figure 7. Average length in words of brackets in English, German and Swedish originals.

The lower proportion of clauses in German originals thus appears to be reflected in shorter brackets compared to English and Swedish. The different writing conventions in the three

[9] English $\chi 2=88.30$, p=<.0001; German $\chi 2=28.41$, p=<.0001; Swedish $\chi 2=41.86$, p=<.0001

languages, which were discussed in Section 2, probably affect these frequencies. We therefore compared the proportions of words occurring in brackets in the three subcorpora. This comparison indicates that English indeed differs from the other languages: 1.4% of all English words appear in brackets, compared to only 0.6% in German and 0.5% in Swedish.

In this section, the findings from the LEGS originals have indicated that brackets are most common in original English and the least common in Swedish. The classification into functional categories suggests that most brackets are content-oriented, and that the unexpected difference between the subcorpora partly originates in content-oriented brackets being even more frequent in English than in the other languages and partly in slightly different compositions of the subcorpora. Furthermore, German texts have the strongest preference for phrasal constructions, a tendency also observed in previous studies (e.g., Carlsson, 2004; Becher, 2011; Ström Herold and Levin, forthcoming). Section 4.2 discusses the findings for the translations in LEGS.

4.2 Brackets in translations

4.2.1 *Congruent and non-congruent translations of brackets*

This section first presents the distributions of brackets in translations compared to originals, then the proportions of brackets retained in translations and finally the frequencies of the different non-congruent translation strategies. To begin with, Figure 8 presents the bracket frequencies in the three original corpora and the six translations in order to determine to what extent they differ.

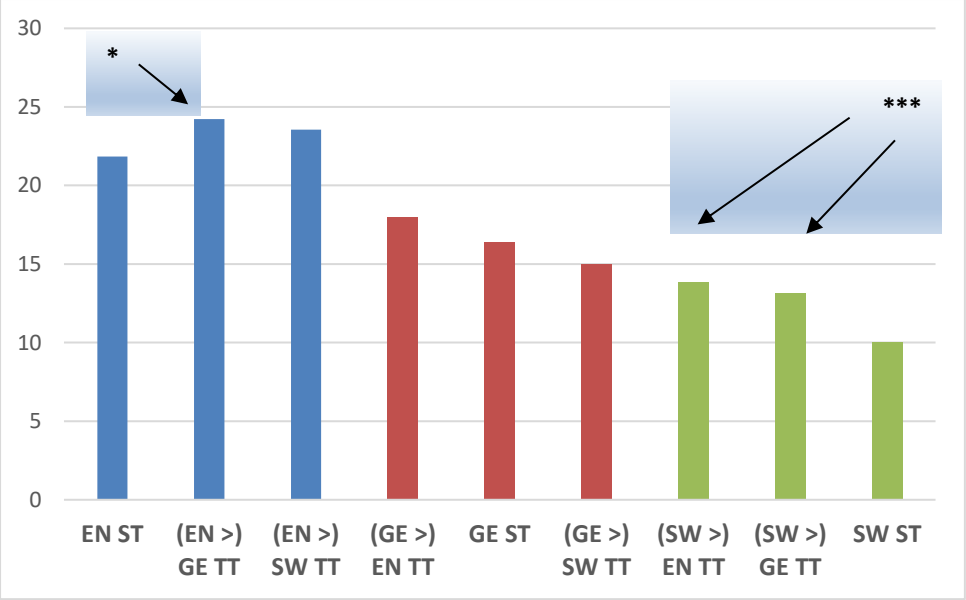

Figure 8. Frequencies per 10,000 words of brackets in originals and translations.

Two trends emerge, but they are difficult to disentangle: firstly, translations tend to contain more brackets than originals and, secondly, translations tend to approach target-language norms rather than adhere to source-text usage. For the first trend, five of six translation subcorpora contain more brackets than their originals (the exception being Swedish translations from German); while, for the second trend, four out of six translations "move towards" target-language norms (the two translations from English originals do not follow this pattern). The three significant differences identified in Figure 8 all relate to higher bracket frequencies in translations. Two of these – the English and German translations from Swedish – produce the

highest significance values. Notably, it is for these where the trends of using more brackets in translation and moving towards target-language norms strive in the same direction.[10] For their English-to-German material, Baumgarten *et al*. (2008: 191−192) conclude that most changes in the German translations are due to adaptation, where translators adapt to the textual conventions of the target language, and not translational explicitation.

In the LEGS material, we identified five different translation strategies applied by the translators: brackets in originals can be I) retained in translations, II) they can be added, III) downgraded, IV) omitted or V) upgraded. The retention strategy simply means that the original brackets are kept in the translation (i.e. congruent translations) without substantial changes to the wording, as in (15) below. Added brackets refer to cases where the translator adds new information in brackets that is not available in the original. In (16), the imperial unit *311 ounces* is added in the translation. In downgrades, a non-bracketed clause or phrase in the original is bracketed in the translation. This is exemplified in (17), where the original Swedish phrase appears between commas but is bracketed in the English translation. Omissions involve instances where the original brackets and the bracketed content are removed, as in (18). Finally, upgrades are the opposite of downgrades. In translation upgrades, the translators remove the brackets while keeping the content. Removing brackets may lead to zero punctuation or, as in (19), the use of another punctuation mark such as dashes.

I Retained

(15) … they were made even more homesick by the horrors of British cuisine, from over-cooked mutton and cabbage to the ubiquitous custard *(which also appalled the Free French)*. [LEGS; English original]
Ihr Heimweh wurde verstärkt durch die Schrecken der britischen Küche, von zerkochtem Lammfleisch mit Kohl bis zu der allgegenwärtigen Vanillesoße *(die auch die Freien Franzosen abstoßend fanden*). (‘which also the Free French found repulsive’) [German translation]

II Addition

(16) Wir verlieren rund *100 Milliliter* Flüssigkeit täglich. [LEGS; German original)]
We lose around *311 ounces (100 milliliters)* of fluid a day. [English translation]

III Downgrade

(17) Ungefär tjugo kort, *alltså tio par*, blandas … (‘i.e. ten pairs’) [(LEGS; Swedish original]
Twenty or so such cards *(i.e., ten or so pairs)* are shuffled … [English translation]

IV Omission

(18) Was die drei Männer verbindet, ist ein Trugschluss: der Regression-zur-Mitte-Irrtum *(englisch: regression toward the mean)*. [LEGS; German original]
What links the three men is a fallacy: the regression-to-mean delusion Ø. [English translation]

V Upgrade

(19) … wenn Ihr Kätzchen das *(höchstwahrscheinlich)* schon nicht tut. (‘most likely’) [LEGS; German original]
… even though your kitten is – *most likely* – not going to share that feeling. [English translation]

[10] Log likelihood English to German p=<.05; Swedish to English p=<.0001; Swedish to German p=<.0001

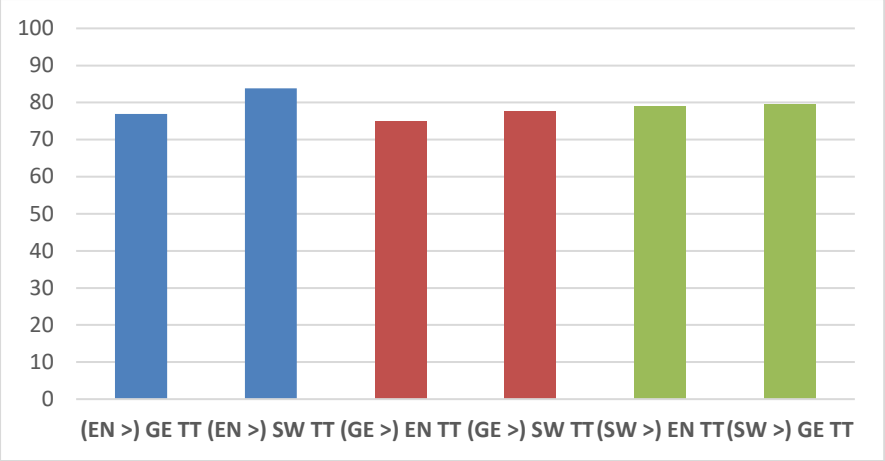

Figure 9. Proportions of retained brackets in translations.

Figure 9 presents the percentages of retained brackets in the translations. Two trends are evident in the LEGS data: Most brackets are retained in translations, and there are systematic target-language-specific preferences for the retention rates. The most obvious result is that a large majority of all brackets are directly transferred in the translations. In all six languages pairs, the retention rate reaches between 75% and 85%. This confirms the strong tendency for direct transfer found previously for the translation of punctuation marks (Gustafsson, 2013; Wollin, 2018; Frankenberg-Garcia, 2019: 23; Ström Herold and Levin, forthcoming). These retention rates nevertheless stand in stark contrast to the low values presented by Baumgarten *et al*. (2008). In their study, only 30% of the brackets were directly transferred from English to German, a finding that is most likely due to specific genre-conventions in their narrowly sampled material. The second trend in our material is that pairwise comparisons between the target languages suggest that Swedish translators retain the most brackets and English translators the least. Figure 10 below sheds further light on this phenomenon by comparing the non-congruent translation strategies in the six target-text subcorpora. Of the translation strategies, additions are the most frequent for five of six translations (the only exception being the German-to-Swedish subcorpus).

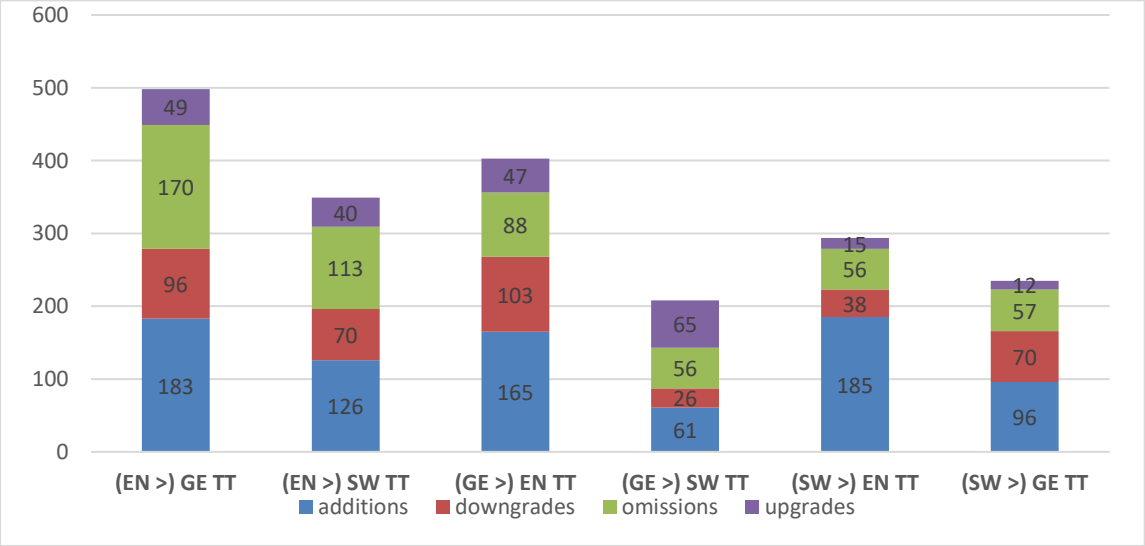

Figure 10. Raw numbers for non-congruent translation strategies per source language.

The proportions of retained brackets in the six translations in Figure 9 are in a complementary relationship with those of the non-congruent translation strategies (additions, downgrades, omissions and upgrades) made in translations in Figure 10. English retains the least brackets, and also uses the most non-congruent strategies, while the language that retains the most brackets, Swedish, uses the least number of non-congruent strategies.

Pairwise comparisons between the target languages show that the non-congruent translation strategies are the mirror image of the retained brackets in Figure 9. Taken together, Figures 9 and 10 thus suggest that English translators retain the least and change the most brackets, Swedish translators retain the most and change the least with German translators consistently between these two. It is difficult to determine exactly why German translators seem more prone to changing punctuation than Swedish translators. One possible explanation is that Swedish, compared to German, is a minor language, and, thus, language-status-wise is lower on the hierarchy. This status difference between the languages is illustrated by the UNESCO's *Index Translationum* where English is the most frequent source language, German the third and Swedish the seventh. Another possible explanation, most likely connected to the afore-mentioned hierarchy, is that that the editors' briefs to the translators may differ depending on target language in that Swedish translators are given less options to interfere.

The next section explores the correlations between non-congruent translations and functional categories, the question being if there are any functions of brackets that are more commonly added or omitted, or otherwise altered in translations.

4.2.2 *Functional categories in non-congruent translations*

This section compares the frequencies of the five functional categories presented and exemplified in Section 4.1 (specification, synonym, reader address, hedge and subjective author comment) and the four non-congruent translation strategies (addition, omission, downgrade and upgrade). As seen in Figure 10, the LEGS data indicate that when translators introduce changes related to brackets, it is most likely to occur in connection with the addition of new information in order to make target texts more explicit to readers.

The 816 additions in the LEGS translations in Figure 11 present a homogenous picture of the distributions across the six language pairs:

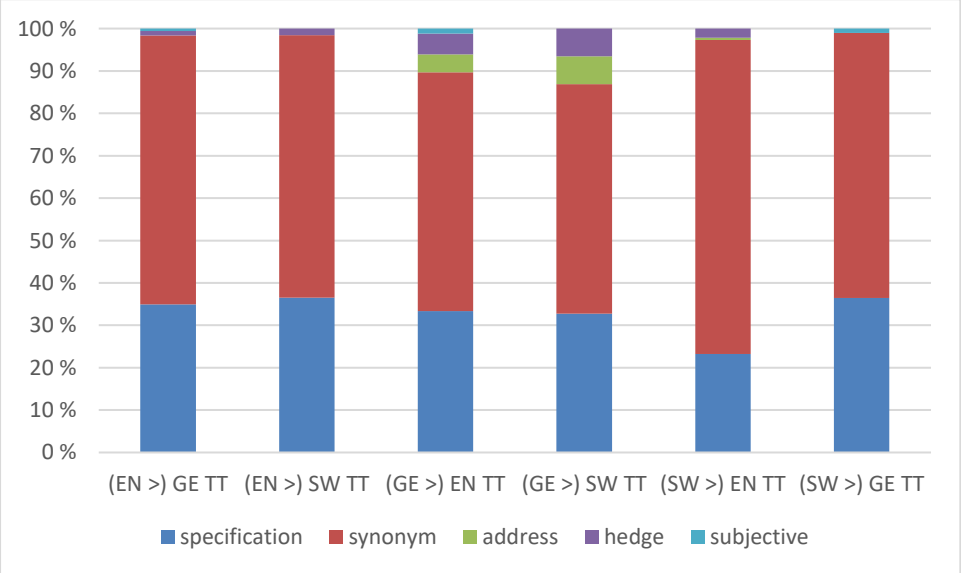

Figure 11. Functional categories of added brackets in translations.

In the majority of cases, additions involve the introduction of synonyms, followed by specifications. Interpersonal brackets are added only rarely (4.4% (36/816)). These trends prevail in all subcorpora. Adding co-referential synonyms, which is also the original function proposed for bracketed text (Biber and Gray, 2016: 205−206), is an unobtrusive way for translators to make the target-text more comprehensible and explicit to the target-text readers. The added synonyms in translations fulfil similar functions to those found in originals. For example, acronyms that are less known in the target-language cultures are spelled out (*the RSPB* > *der RSPB (Königliche Gesellschaft für Vogelschutz)* (Ge.); *RSPB (Kungliga fågelskyddssällskapet)* (Sw.)), additional name variants are given (*bei Klausenburg* (Ge.) > *near Kolozsvár (Cluj)*; *vid Cluj (Klausenburg)* (Sw.)) and sometimes phrases are also given in the target language (*BP rebranded itself "Beyond Petroleum"* > *definierte BP sein Kürzel von „British Petroleum" in „Beyond Petroleum" (Jenseits von Erdöl) um* (Ge.)). Added specifications mainly serve the purpose to explicitate various cultural phenomena that are not likely to be well known to target-text readers (*Engelbrektsmarschen* (Sw.) > *the Engelbrekt March (named after the fifteenth-century rebel leader and proto-nationalist Engelbrekt Engelbrektsson*)).

The second largest category of the non-congruent strategies, the 540 omissions, is more varied. As seen in Figure 12, the translations from English differ from those from German and Swedish:

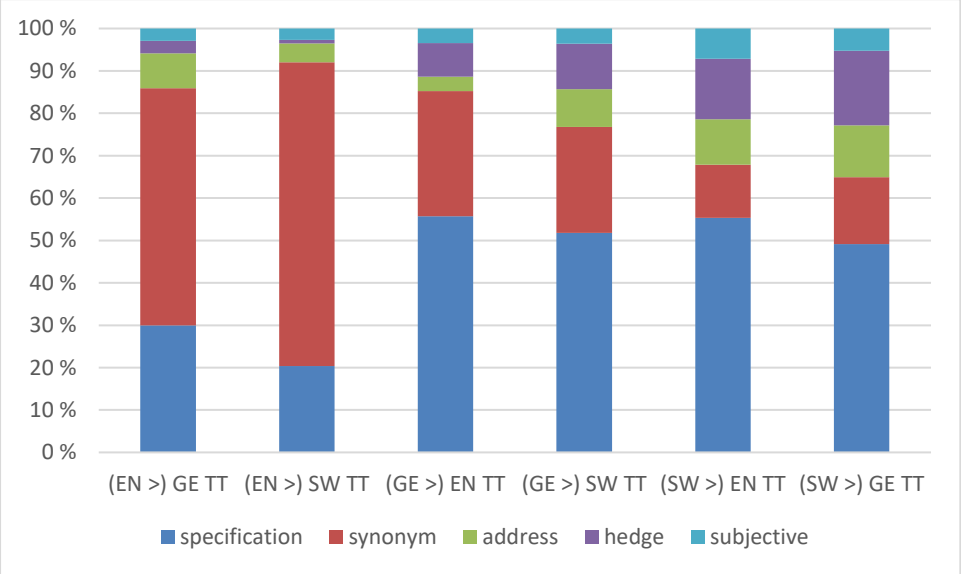

Figure 12. Functional categories of omitted brackets in translations.

The German and Swedish translations from English consistently omit the imperial units (e.g., *three thousand meters (ten thousand feet)* > *3000 Meter* (Ge.); *3 000 meter* (Sw.)), which partly explains the high frequencies of omitted synonyms. Other instances of omitted synonyms relate to the source language, in contrast to the target language(s), sometimes using more than one term for a concept, e.g., a Greek or Latin term and a native term (*Chlorophyll (Blattgrün)* (Ge.) > *chlorofyll*; *klorofyll*) or the source text itself containing a translation couplet that appears superfluous in the target text (*With typical Nazi bombast […] codenamed Adlerangriff (Eagle Attack)* > *erhielt […] den bombastischen Codenamen „Adlerangriff"* (Ge.)). In some cases, translators omit bibliographical or historical data, as when the English *Charles Darwin (1809–82)* is rendered only as *Charles Darwin* in the German translation. Interestingly, Baumgarten *et al*. (2008: 190) seem to indicate that this kind of information instead tends to be added rather than omitted in their English-to-German translations. Further studies are needed to determine

if any specific kind of information would be more or less likely to be added or omitted cross-linguistically.

Other omissions in the LEGS data concern specifications manifested as source-text elaborations on terms (see example (20)) or additional source-text information on culture-specific items (as in (21)) deemed superfluous by translators.

(20) Attacks mainly by Stuka dive-bombers and by *fast S-Boote (motor torpedo boats which the British called E-boats)* virtually closed the Channel to British convoys. [LEGS; English original]
Anfall av främst Stuka-störtbombare och *snabba motortorpedbåtar* praktiskt taget stängde Engelska kanalen för de brittiska konvojerna. [Swedish translation]

(21) … men när man ville klä majstång visade det sig svårt att vid den här årstiden *(som dessutom var förskjuten framåt ett par veckor på grund av vår på den tiden omoderna kalender)* få tag på blommor och grönt. (‘(which besides was moved forward two weeks because of our at that time outdated calendar)’) [LEGS; Swedish original]
… however when it came time to decorate the May pole it proved difficult to find enough flowers and greenery at that time of year. [English translation]

As for most omissions, Figures 13 and 14 show that a majority of all downgrades and upgrades involve specifications in all subcorpora. Numbers are lower in these categories (403 downgrades and 228 upgrades).

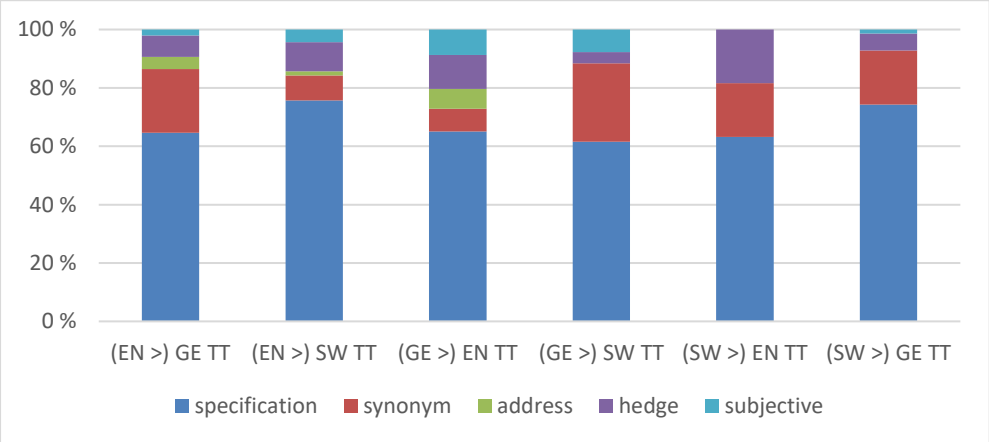

Figure 13. Functional categories of downgraded brackets in translations.

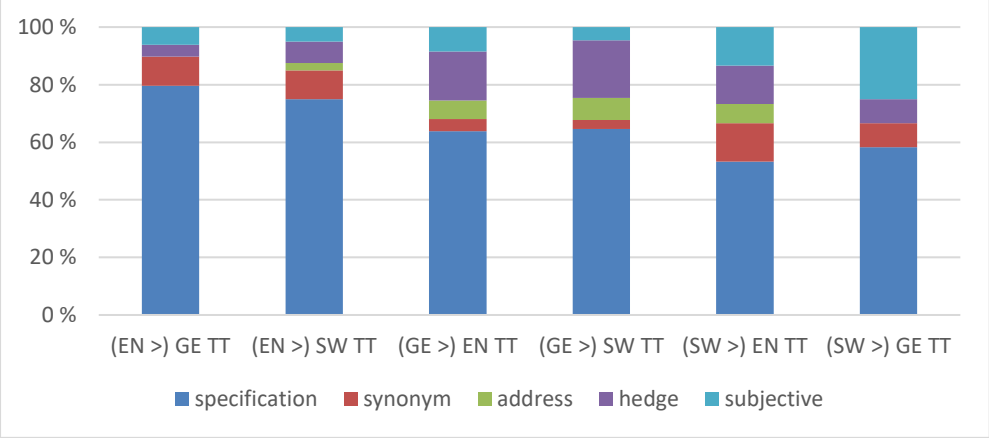

Figure 14. Functional categories of upgraded brackets in translations.

The choices in individual instances are difficult to explain in relation to the categories of brackets. Specifications account for a majority of all tokens in the three original corpora and this is also the case for both downgrades and upgrades. Individual translators' choices largely seem to determine the usage. For instance, in (22), the English translator chooses to downgrade the German original phrase occurring between dashes, while the Swedish translator, as seen in (23), upgrades the bracketed English original clause by putting it between commas (for further discussions of other punctuation marks as correspondences of brackets, see Section 4.2.3). In (22), the downgrade is introduced in English, the language that is most likely to use non-congruent translations. In (23), on the other hand, the upgrade is introduced in Swedish, which is the language least likely to change punctuation in translation. (For more on punctuation marks in downgrades and upgrades, see next section.)

(22) … in den digitalen Ordnern, die wir nun Tag für Tag – *und immer wieder auch Abend für Abend* – durchsehen, … ('and ever again also evening for evening') [LEGS; German original]
… in the computer folders we mine day after day (*and often night after night*), … [English translation]

(23) … *the Clinton administration (which took office in January 1993)* … [LEGS; English original]
… *Clintonregeringen, som tillträtt i januari 1993,* … ('the Clinton-administration which [had] taken office') [Swedish translation]

What is noteworthy is that upgrades are the smallest category of change in five out of six translations. As with additions, German-to-Swedish is the exception here, and somewhat unexpectedly, this category is even the largest in this subcorpus. The differences between this subcorpus and the others should nevertheless not be overemphasized, since this particular subcorpus has the lowest number of changes and the differences between the categories are fairly small. The next section will further explore downgrades and upgrades regarding the punctuation marks used as correspondences of brackets.

4.2.3 *Corresponding punctuation marks in downgrades and upgrades*

This section presents a different perspective on translation changes by investigating the punctuation marks that correspond to brackets in downgrades and upgrades. The findings are shown in Figures 15 and 16. As is evident, commas and no punctuation marks ('zero') are the most frequent correspondences. Although dashes are often mentioned as being used in similar functions as brackets, they account for only 10% of all correspondences in the figures. It is likely that the informal connotations of dashes mentioned in Section 3 (Quirk *et al*., 1985: 1629; Leech *et al*., 2009: 245; Crystal, 2015: 158) make them less likely as correspondences of brackets, which, according to Leech *et al*. (2009: 246), are more typical of "serious written style".

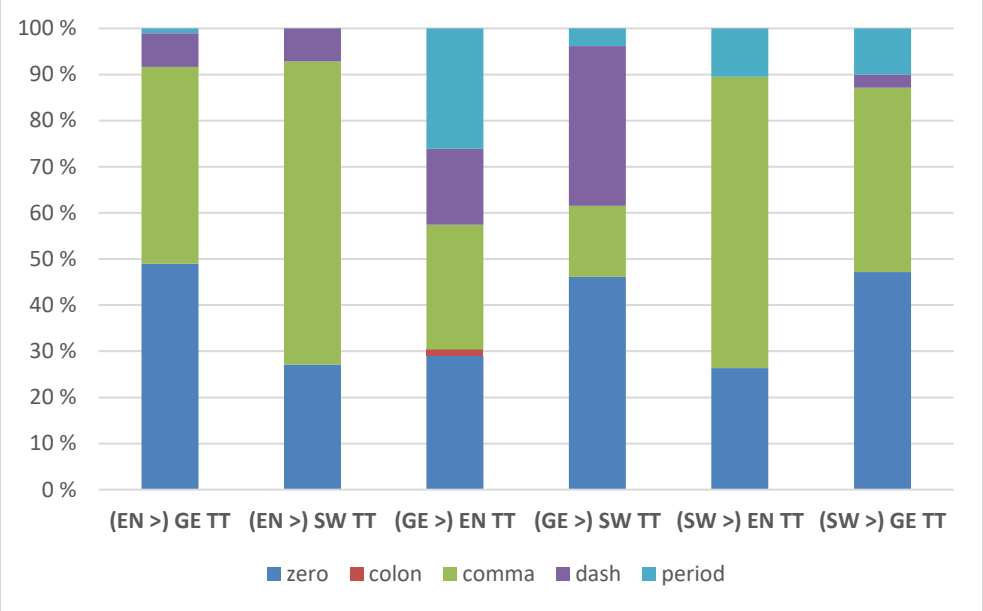

Figure 15. Source-text punctuation in downgrades.

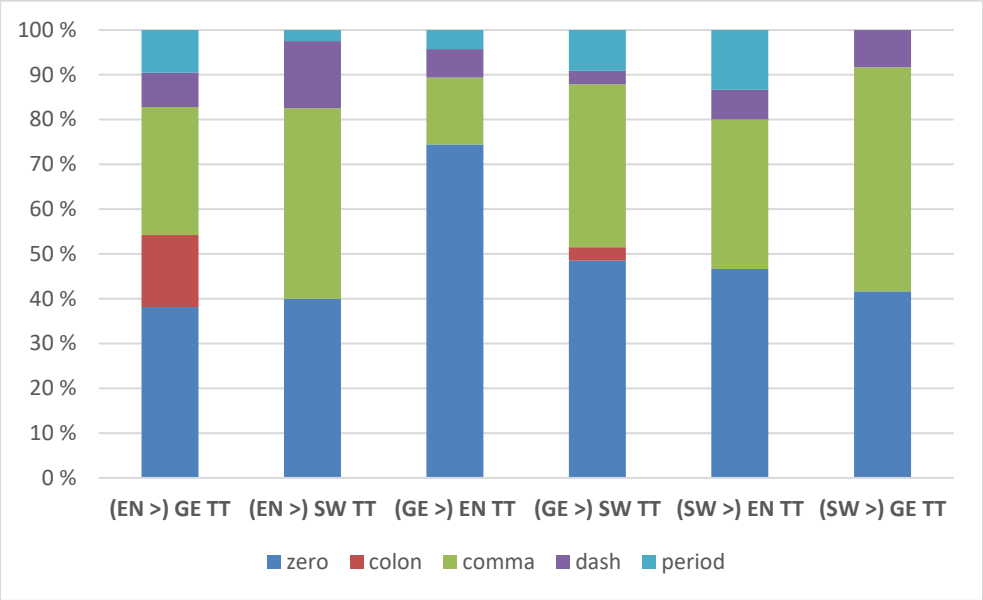

Figure 16. Target-text punctuation in upgrades.

In all translation directions, and this holds for both downgrades and upgrades, either zero punctuation or commas are the most common correspondences. In all subcorpora, these two alternatives account for more than half the instances. In fact, only in the downgrades from German originals do zero punctuation and commas combined not exceed 70% of the tokens. It should be noted that zero marking and commas were also the most common (non-colon) correspondences of colons in Ström Herold and Levin (forthcoming). It would therefore seem that the frequencies and flexibility of these two options make it more likely for them to be changed from or into brackets in translation. An additional important factor is that commas, like brackets, are correlative marks having similar stylistic values. However, as pointed out by *English Style Guide* (2016/2019: 12), replacing brackets with commas may have information-structural consequences regarding the degree of backgroundedness, i.e., bracketed material being more strongly backgrounded than material placed between commas. Translators are more prone to replacing commas with brackets than the other way around. As brackets are considered

to have a communicative function (Bredel, 2018: 12f.), the introduction of brackets in the translation entails a stronger presence of the writer. However, in translations, readers can usually not distinguish between the authors' and the translators' contributions[11] (e.g., *his parents' visits to the nursery, usually at tea-time on Sundays, could be excruciating occasions* > *föräldrarnas besök i barnkammaren (vanligtvis vid tedags på söndagarna) kunde vara mycket plågsamma* (Sw.) ['usually at tea time on Sundays']). Thus, translations are in this respect different from originals, where it is obvious that all brackets stem from the author.

There are some minor tendencies observable for punctuation changes in downgrades. In some instances, one member of a source-text apposition is put in brackets in the translation (*the giant sloth megatherium* > *das Riesenfaultier (Megatherium)* (Ge.)). In others, the target text disallows a certain word order, which leads to obligatory restructuring. For instance, the year in the original English *JOHN CHEEVER'S 1961 SHORT STORY "The Angel of the Bridge"* must be moved because German does not normally compound years with nouns (**1961-Kurzgeschichte*). Instead, the year appears after the title in brackets, as is conventionally done in many languages: *JOHN CHEEVERS KURZGESCHICHTE „Der Engel der Brücke" (1961)*. In such instances, the change of punctuation mark is a way for translators to solve structural translation problems.

For upgrades, we find similar punctuation changes, though in the opposite direction to downgrades. Thus, a German original introducing an acronym in brackets, *von Massenmorden der sowjetischen Geheimpolizei (NKWD)* ['mass murders of the Soviet secret police (NKVD)'], was rendered into Swedish as a non-bracketed apposition, *den sovjetiska hemliga polisen NKVD:s massmord*. Similarly, English translators condense noun phrases by moving the originally bracketed year into premodifying position (*wie im Song Going Back von Carole King (1966)* ['like in the song'] (Ge.) > *think of Carole King's 1966 song Going Back*). As shown by Ström Herold and Levin (2019), English noun premodifiers often have different kinds of translation correspondences in German and Swedish, such as postmodifiers or genitives.

Thus, as seen in Sections 4.2.2 and 4.2.3, downgrades and upgrades are rather rare in our material and major generalizations are not easily made. One finding is nevertheless that downgrades and upgrades are often reflections of each other in that a structure that is downgraded into brackets in one text may typically be the result of an upgrade in another. In both downgrades and upgrades, commas and zero punctuation are the most frequent correspondences. The final section shifts focus from punctuation correspondences in downgrades and upgrades to structural changes to bracketed text that is retained in translations.

4.2.4 *Clause building and clause reduction in translations*

The distributions of clauses and phrases in original brackets in Figure 6 above show that there is a stronger preference in German than English and Swedish for putting phrases in brackets. We therefore decided to also compare the distributions of clauses and phrases in translations with their originals. This comparison was restricted to retained brackets, i.e., added, omitted, downgraded and upgraded brackets were not included. In this manner, it is possible to compare the proportions of original clauses and phrases rendered as either clauses or phrases in translations.

Previous studies (Dirdal, 2014; Ström Herold and Levin, 2018, forthcoming) have observed that translations tend to use more elaborate structures than originals, meaning that translators tend to go from phrasal to clausal constructions rather than the other way around. Dirdal (2014: 122) uses the term 'clause building' for those changes that move towards independent main clauses, such as phrases rendered as clauses or non-finite clauses rendered

[11] Unless the brackets are explicitly labelled "translator's note".

as finite clauses. The term 'clause reduction' is used to refer to the opposite case. Compared to clause building, clause reduction is usually rare (Dirdal, 2014; Ström Herold and Levin, forthcoming). The tendency towards clause building is also present in our material, but, as will become evident below, only partially. Contrary to Dirdal (2014), we restricted our study to a binary opposition between clausal instances, which consist of verbs in combination with other elements, and phrasal instances, which do not, and instead typically contain noun phrases or prepositional phrases.

In our data, it is only the German originals that trigger significant proportions of clause building. English and Swedish originals, which contain significantly more clauses than German originals (as shown in Figure 6 above), do not yield significant differences between originals and translations. Figure 17 below therefore focuses only on German originals and their translations. The four bars illustrate the distributions of phrases and clauses in the two translations with the blue colour indicating German original phrases and the orange indicating original clauses.

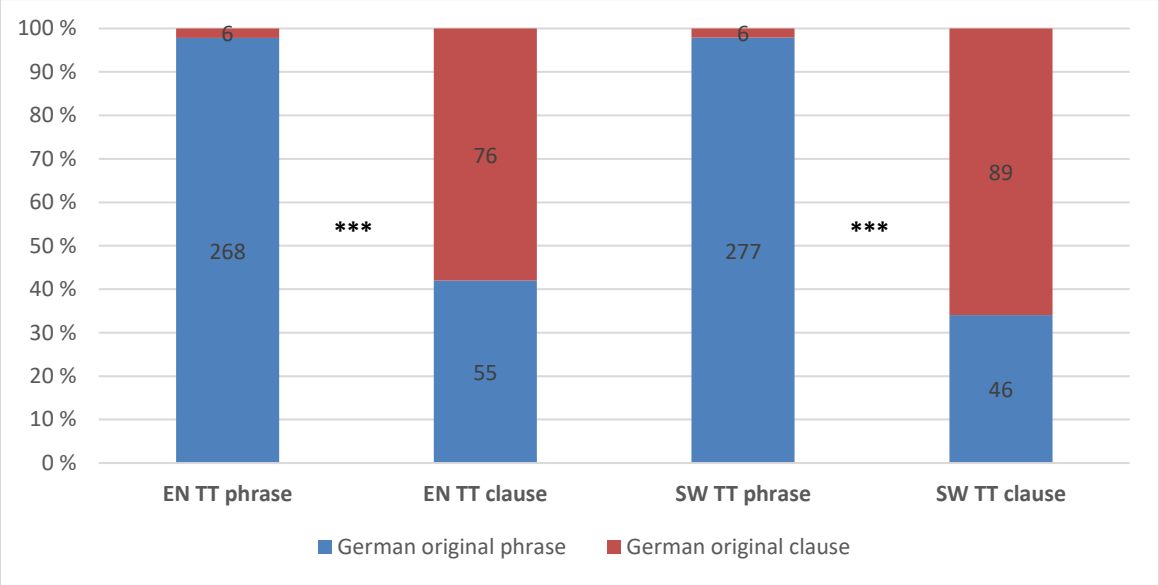

Figure 17. Clause building and clause reduction in translations from German.

The first and third bars in the figure show that German original clauses are very rarely translated into English or Swedish phrases, while phrase-to-phrase translation is the norm. The second and fourth bars show that fairly large proportions of the target-text clauses originate in German source-text phrases. There is a significant likelihood for more phrases to be rendered as clauses than for clauses to be rendered as phrases from these source texts.

Examples (24) and (25) illustrate clause building in the translation of the bracketed text. The German original in (24) explains the term *Mikrobiota* in a condensed manner using a descriptive noun phrase which, interestingly, is introduced by an equal sign. This explanation is then structurally expanded into an English relative clause in the translation.

(24) Wissenschaftlich korrekt sagt man Mikrobiota *(= kleine Leben)* oder auch Mikrobiom … ('= small living beings') [LEGS; German original]
The scientifically correct terms are microbiota *(which means "little life")* and microbiome, … [English translation]

The next example, contrary to the short, content-based one above, fulfils an interpersonal function and involves more lexical components expressing a nuanced hedged comment. The

fragmentary German style of the original is avoided by both translators who instead use full finite clauses by adding subjects, verbs and objects.

(25) Einmal sagte er *(möglicherweise sogar in vollem Ernst)*, der beste Augenblick seiner Präsidentschaft sei gewesen, als … (‘possibly even in all seriousness’) [LEGS; German original]
Once he said *(and he may have meant it seriously)* that the finest moment of his presidency was when … [English translation]
En gång sa han *(och kanske menade han rentav allvar)* att det bästa ögonblicket under hans presidenttid var då … (‘and maybe he meant it seriously’) [Swedish translation]

Our German original data thus support previous findings on clause building and clause reduction in that translators at least from this source language tend to build clauses rather than reduce clauses to phrases.

5. Conclusions

Our investigation into brackets in original texts and translations in nine subcorpora of LEGS has shed light on both similarities and notable differences. In originals, English stands out as the most bracket-friendly language. Brackets are less common in German originals and the rarest in Swedish. However, further studies are needed to establish to what extent these findings are generalizable to other subgenres than those included in our study, such as newspaper text. As for the function of bracketed texts, the languages behave similarly: most brackets are content-oriented as opposed to interpersonal, which suggests that brackets largely have kept their original information-condensing function (cf. Biber and Gray, 2016: 205−206). Still, the relatively large proportions of interpersonal brackets and the differences between our operative and more content-oriented texts indicate that there is an ongoing expansion of functions and frequencies in all three languages.

In translations, most brackets are retained, indicating a fairly high degree of source-text adherence which aligns well with previous translation studies on punctuation. If retention is the most frequently used option, the addition of brackets in translation is the most common non-congruent strategy. Added brackets mostly consist of short synonyms, which suggests that brackets are a frequent unobtrusive means for translators to move the text to the target readers by, e.g., spelling out acronyms or giving additional name variants.

Our results are inconclusive regarding which functional aim is the strongest: do translations usually contain more brackets than their originals because translators strive to bring the text closer to the target-text readers or do translators aim for the target-text norms? Further studies are certainly warranted, though an interesting finding is that the most marked differences between original and translations appears with Swedish originals, where the two tendencies would promote increased use of brackets.

The present findings are similar to those found for colons (Ström Herold and Levin, forthcoming) in that English translators are more “daring”, using the most non-congruent translations. Swedish translators introduce the least and this difference may be connected to differences in language status and power relations. Another finding already indicated in previous studies relates to the nominal style in German (Carlsson, 2004). In view of this, it is not surprising that English translators often build clauses (cf. Dirdal, 2014) from phrases in translations from German. The tendency for clause building is nevertheless strong enough for Swedish translators to also build clauses in translations from German, in spite of their generally more cautious approach to the introduction of changes.

By using a bidirectional trilingual translation corpus, we have, at least to some extent, been able to tease apart specific language preferences and translation-induced changes. However, as with most studies, the present study calls for further investigations. Will similar results be found with different language pairs, genres and punctuation marks? The competition between translation-specific and language-specific tendencies that may or may not work against each other will certainly constitute a fruitful field for future work.

References

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

Baker, M. 1996. Corpus-based Translation Studies: The Challenges that Lie Ahead. In *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, H. Somers (ed.), 175–186. Amsterdam: John Benjamins.

Baumgarten, N., Meyer, B. and Özçetin, D. 2008. Explicitness in translation and interpreting: A critical review and some empirical evidence (of an elusive concept). *Across Languages and Cultures* 9(2): 177–203.

Becher, V. 2011. Von der Hypotaxe zur Parataxe: Ein Wandel im Ausdruck von Konzessivität in neueren populärwissenschaftlichen Texten. In *Satzverknüpfungen. Zur Interaktion von Form, Bedeutung und Diskursfunktion*, E. Breindl, Gi. Ferraresi and A. Volodina (eds), 181–209. Berlin: de Gruyter.

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

Bisiada, M. 2013. Changing Conventions in German Causal Clause Complexes. A Diachronic Corpus Study of Translated and Non-translated Business Articles. *Languages in Contrast* 13(1): 1–27.

Bredel, U. 2018. Das Interpunktionssystem des Deutschen. *Studia Neophilologica* 90(1): 7–23.

Carlsson, M. 2004. *Deutsch und Schwedisch im Kontrast: Zur Distribution nominaler und verbaler Ausdrucksweise in Zeitungstexten*. PhD dissertation, Gothenburg University.

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

Dirdal, H. 2014. Individual Variation between Translators in the Use of Clause Building and Clause Reduction. In *Corpus-based Studies in Contrastive Linguistics*, S.O. Ebeling, A. Grønn, K.R. Hauge and D. Santos (eds), *Oslo Studies in Language* 6 (1): 119–142.

Duden. Band 9. Das Wörterbuch der sprachlichen Zweifelsfälle. Richtiges und gutes Deutsch. 8th ed. 2016. Berlin: Dudenverlag.

English Style Guide. A handbook for authors and translators in the European Commission. 2016/2019. European Union.

Englund Dimitrova, B. 2014. Till punkt och pricka? Översättarstil, normer och interpunktion vid översättning från bulgariska till svenska. *Slovo. Journal of Slavic Languages, Literatures and Cultures* 55: 77–99.

Frankenberg-Garcia, A. 2019. A Corpus Study of Splitting and Joining Sentences in Translation. *Corpora* 14(1): 1–30.

Gustafsson, R. 2013. *Att översätta kolon. En undersökning av hur skiljetecknet kolon överförs från franska till svenska i skönlitterära översättningar.* MA thesis. Gothenburg University.

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

House, J. 2011. Using Translation and Parallel Text Corpora to Investigate the Influence of Global English on Textual Norms in Other Languages. In *Corpus-based translation studies*, A. Kruger, K. Wallmach and J. Munday (eds), 187–208. London: Bloomsbury.

Index Translationum. UNESCO. N.d. Available at: http://www.unesco.org/xtrans/ [Last accessed 4 June 2021]

Ingo, R. 2007. *Konsten att översätta. Översättandets praktik och didaktik*. Lund: Studentlitteratur.

Kranich, S. 2011. To Hedge or Not to Hedge: The Use of Epistemic Modal Expressions in Popular Science in English Texts, English–German Translations, and German Original Texts. *Text & Talk* 31: 77–99.

Leech, G., Hundt, M., Mair, C. and Smith, N. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge: Cambridge University Press.

Nádvorníková, O. 2020. The Use of English, Czech and French Punctuation Marks in Reference, Parallel and Comparable Web Corpora: A Question of Methodology. *Linguistica Pragensia* 30(1): 30–50.

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

Shiyab, Said M. 2017. *Translation: Concepts and Critical Issues*. Antwerp: Garant Publishers.

Ström Herold, J. and Levin, M. 2018. English supplementive *ing*-clauses and their German and Swedish correspondences. In *Corpora et Comparatio Linguarum: Textual and Contextual Perspectives*, S.O. Ebeling and H. Hasselgård (eds.), *Bergen Language and Linguistics Studies* 9(1), 115–138.

Ström Herold, J. and Levin, M. 2019. *The Obama Presidency*, *the Macintosh Keyboard* and *the Norway Fiasco*: English Proper Noun Modifiers in German and Swedish Contrast. *English Language and Linguistics* 23(4): 827–854.

Ström Herold, J. and Levin, M. Forthcoming. The Colon in English, German and Swedish: A contrastive corpus-based study. *Comparative Punctuation*. Linguistik – Impulse & Tendenzen. Berlin, New York: Walter de Gruyter.

Wollin, L. 2018. Punctuation: Providing the Setting for Translation? *Studia Neophilologica* 90(1): 37–49.

Authors' addresses

Magnus Levin / Jenny Ström Herold
Linnaeus University
Department of Languages
SE-351 95 Växjö
Sweden
magnus.levin@lnu.se / jenny.strom.herold@lnu.se

Lexicogrammar through colligation: Noun + Preposition in English and Norwegian

Hilde Hasselgård

University of Oslo (Norway)

This study compares sequences of noun and preposition in English and Norwegian using data from the English-Norwegian Parallel Corpus. One purpose is to test the use of sequences of part-of-speech tags as a search method for contrastive studies. The other is to investigate the functions and meanings of prepositional phrases in the position after a noun across the two languages. The comparison of original texts shows that the function of postmodifier is most frequent in both languages, with adverbial in second place. Other functions are rare. English has more postmodifiers and fewer adverbials than Norwegian. Furthermore, the prepositional phrases express locative meaning, in both functions, more frequently in Norwegian than in English. The study of translations reveals that the adverbials have congruent correspondences more often than postmodifiers, particularly in translations from English into Norwegian.

Keywords: prepositional phrase, colligation, postmodification, adverbial, English/Norwegian

1. Introduction

The contrastive study of lexicogrammar involves the challenge of identifying search strings that can retrieve the same type of construction(s) in both languages investigated. This challenge has most frequently been met by identifying a lexical correlate of particular constructions (Johansson, 2007: 37). This study uses a sequence of part-of-speech (PoS) tags as a window into cross-linguistic syntactic differences and similarities. The selected tag sequence is noun plus preposition, expected to retrieve nouns with a postmodifying prepositional phrase (PP) as well as chance sequences of a noun and a PP with adverbial function, as illustrated by (1) and (2) from the English-Norwegian Parallel Corpus (ENPC) on which this study is based. Both examples have congruent translations, suggesting structural similarity between English and Norwegian PPs.[1]

[1] Corpus examples are written as they occur in the corpus, with the source text first. Any abbreviations are marked with three dots. Identification tags ending in T (e.g. HW2T) indicate that the example is a translation. Norwegian examples are followed by a transliteration marked "Lit." to clarify the structure, except where the published translation is word-for-word equivalent.

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.v11i1.3442.

(1) Nor did he enjoy his *meetings with* Dr Forestier… (BC1)
Han likte heller ikke *konsultasjonene hos* doktor Forestier … (BC1T)
Lit: "He liked either not the consultations at doctor Forestier"

(2) Og han hadde lagt *klærne på* en stein mye lenger opp. (HW2)
Lit: "And he had laid the clothes on a rock much further up."
And he had laid his *clothes on* a rock much nearer the grove. (HW2T)

The noun in the sequence provides a grammatical context for the PP. In the case of postmodifying PPs it will typically be the head of a complex noun phrase, and it will be relevant to investigate the meaning relation between the head and the PP. PPs functioning as adverbials, however, are not part of the same syntagm as the preceding noun, so that the noun and the PP will be more peripherally related semantically too, if at all.

The following research questions are addressed:

- What are the syntactic functions of PPs following a noun in Norwegian and English?
- What meanings do the PPs convey?
- Are there quantitative and qualitative differences between the languages as regards the functions and meanings of postnominal PPs?
- To what extent are translations congruent?

The use of PoS tag sequences as a starting point has not been common in cross-linguistic corpus studies (though see Wilhelmsen, 2019 and monolingual studies of L1 and L2 performance, e.g. Granger and Rayson, 1998; Granger and Bestgen, 2014). Hasselgård (2016) searched for a combination of function words and wildcards ('the * of the *') as a colligational framework (Renouf and Sinclair, 1991) for a contrastive study of complex noun phrases. A major weakness of this search method was that it was impossible to identify an equivalent colligational framework for Norwegian, so that Norwegian was studied only through the English pattern (Hasselgård, 2016: 77). The search method used in the current study is one that should have equal potential in both languages and furthermore casts the net wider to include more prepositions. As shown in examples (1) and (2), it elicits not only complex noun phrases but also sequences with other functions. Hence, the tag sequence 'noun + preposition' should be able to illuminate cross-linguistic syntactic differences between English and Norwegian to do with both postmodification of nouns and clause-level adverbials. The proportional distribution of these functions may in turn indicate preferences towards a nominal or a clausal style.

English is expected to have more postmodifying PPs and Norwegian to have more adverbial PPs. This is based on the finding that postmodifying *of*-phrases frequently have non-congruent Norwegian correspondences (Hasselgård 2016). Furthermore, the claim that English is more nominal while Norwegian is more verbal/sentential (e.g. Nordrum, 2007; Behrens, 2014), might promote postmodifying PPs in English and clause-level adverbials in Norwegian.

The remainder of the paper is organized as follows: Section 2 summarizes some previous studies of prepositional phrases in English and Norwegian. The material and method of the study are outlined in Section 3. Section 4 presents the classificatory framework before the investigation itself appears in Section 5. Section 6 offers a summary of the findings and some concluding remarks.

2. Prepositional phrases in English and Norwegian

Prepositional phrases are structurally similar in English and Norwegian, except that Norwegian prepositions can be complemented by the equivalents of *to*-infinitives and *that*-clauses (*å*-infinitives, *at*-clauses) (Holmes and Enger, 2018: 323). In both languages, prepositions can be stranded after their complement, and occasionally, a preposition is postposed, as in the case of English *ago* (e.g. *three weeks ago*).[2]

In both English and Norwegian, prepositional phrases are common realizations of adverbials and noun postmodifiers (Biber *et al*., 1999: 104; Holmes and Enger, 2018: 401).[3] According to Biber *et al*., PPs are "by far the most common type of postmodification in all registers" (1999: 607) and furthermore the most frequent realization of adverbials, particularly of the circumstantial type (1999: 768; see also Hasselgård, 2010: 38). Similarly, Elsness (2014: 95) shows that prepositions are the most frequent part of speech to follow the tag sequence 'determiner + noun'.

Prepositional phrases also have other syntactic functions. Fang (2000) lists nine in a study of English PPs based on the ICE-GB corpus.[4] Adverbial and NP postmodifier are by far the most frequent ones, accounting for close to 90% of the PPs, but PPs also function as postmodifier of adjectives and adverbs, subject and object complements, complement of preposition, focus of *it*-cleft and stand-alone phrase (Fang, 2000: 188). A similar list of the functions of Norwegian PPs is found in Faarlund *et al*. (1997: 411).

While the 'noun + PP' sequence may superficially resemble a pattern (in the sense of Hunston and Francis, 2000), it is in fact not. More precisely, there may be instances of patterns among the sequences extracted from the corpus, where the preposition is selected by the noun and the PP can be seen as a complementation of the noun (Hunston and Francis, 2000: 40). Such patterns are written either as **N prep** or with a specific preposition such as **N *of* n** (ibid.: 57). If the preposition is not constrained by the preceding noun, there is no pattern even if the 'noun + prep' sequence may be frequent. As Hunston and Francis point out: "frequent co-occurrences of words do not necessarily indicate the presence of a pattern" (2000: 71).

Contrastive studies of the syntactic functions of PPs indicate that languages may have different restrictions and preferences regarding their use even when the linguistic resources are similar. For example, Mott (2013) finds that postmodifying PPs are more restricted in Spanish than in English, particularly in locative expressions (2013: 168), which he links to differences in lexicalization and grammaticalization. Similarly, Moreira-Rodríguez (2006) finds English postmodifying PPs more flexible than Castilian Spanish ones, which may cause English-speaking learners of Spanish to overuse PP modifiers at the cost of relative clauses.

There are not many contrastive studies of PPs in English and Norwegian. In a series of studies, Thomas Egan (e.g. Egan, 2013) discusses the semantics and cross-linguistic correspondences of a number of prepositions, but focuses less on their syntactic functions. As noted above, Hasselgård (2016) compares the English pattern '*the* N1 *of the* N2' to its Norwegian correspondences, noting a high degree of divergence, particularly due to the fact that Norwegian lacks a preposition equivalent to *of*, whose main role is to "[combine] with preceding nouns to produce elaborations of the nominal group" (Sinclair, 1991: 83). Thus, a number of '*the* N1 *of the* N2' sequences correspond to compound nouns, *s*-genitives and expressions involving adverbs (Hasselgård, 2016: 65; see also Holmes and Enger, 2018: 355).

[2] For arguments for the analysis of *ago* as a preposition, see Huddleston and Pullum (2002: 632).

[3] Holmes and Enger (2018) do not present frequency data, but the functions of postmodifier and adverbial are mentioned first under the functions of PPs, possibly indicating an order of importance.

[4] The main objective of Fang (2000) is to test a lexical model for the automatic assignment of syntactic function.

Furthermore, in a study of clausal postmodification of nouns in English and Norwegian, Elsness (2014: 91) argues that "there are some notable differences in the structure of the noun phrase between the two languages", particularly in the use of modifiers. His results suggest that Norwegian prefers more "explicit" noun modification than English, for example favouring finite postmodifying clauses over phrasal postmodifiers, which are considered a more "compact" type of modification (see also Biber and Gray, 2016: 232). Behrens (2014: 157) observes a greater preference for nominalizations in English than in Norwegian academic prose with a correspondingly higher number of actions and events coded as clauses in Norwegian. This will have consequences for the function of associated PPs, which will be postmodifiers in the case of nominalizations and adverbials in the case of clausal expressions.

While the above-mentioned studies have identified some cross-linguistic differences regarding PPs as postmodifiers, there is less reason to expect the same kind of differences in the realization of adverbials. For example, Hasselgård (2021: 211) finds that similar proportions of time adverbials in English and Norwegian news discourse are realized by prepositional phrases. A contrastive analysis of adjunct adverbials in clause-initial position in fiction points in the same direction (Hasselgård, 2014: 85). Previous studies thus suggest that the greatest cross-linguistic differences will be found with PPs functioning as postmodifiers of nouns. Furthermore, it is expected that the languages will differ as to the relative frequencies of what is believed to be the main functions of PPs, namely postmodifiers and adverbials.

3. Material and method

The material for this study comes from the fiction part of the English Norwegian Parallel Corpus (ENPC fiction), which was accessed through the Glossa interface (Johannessen *et al*., 2008). The ENPC is a bidirectional parallel corpus in which originals and translations are aligned at sentence level (Johansson, 2007: 11, 14). The fiction part consists of 30 original text extracts (totalling just above 400,000 words) in either language with translations into the other.[5] In the Glossa version, all original and translated texts are fully PoS-tagged.[6]

The search string entered in the 'Extended search' form in Glossa was 'noun' followed by 'preposition' with no elements allowed in between. This string does not give full recall of postmodifying PPs, since postmodifiers need not follow their head noun directly. In the case of adverbial PPs, the recall is even lower, since adverbial PPs are by no means restricted to postnominal position. However, the postnominal position is one where both functions can be found in PPs, which counts greatly in its favour, considering that the study of alternation is a major concern of this paper. Thus, bearing in mind that the position immediately after a noun is likely to enhance the number of PPs functioning as postmodifiers, the tag sequence 'noun + preposition' was preferred to searches for prepositions only.

The searches were made only in original English and Norwegian texts, but the aligned translations were also retrieved in order to perform the study of translations presented in Section 5.6. The corpus searches returned 17,146 'noun + preposition' sequences in Norwegian and 17,830 in English, corresponding to 4,249 and 4,430 per 100,000 words, respectively. Due to the need for manual analysis of the functions, meanings and correspondences of the prepositional phrases, it was necessary to reduce the material to random samples. The sample size was set to 500 concordance lines per language.

Each concordance line in the random samples was scrutinized to make sure it actually represented a sequence of noun and preposition. This manual sifting revealed some tagging

[5] See https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/sub-corpora/.
[6] The English texts were tagged with the TreeTagger (www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) and the Norwegian texts with the Oslo Bergen Tagger (Johannessen *et al*., 2012).

errors. Concordance lines were excluded if the noun tag was wrong (as in example (3), where *gossip* is a verb) or if the preposition tag was wrongly assigned, as in (4), where the highlighted word is a relative pronoun – a type of word which is never, to my knowledge, classified as a preposition.

(3) He said, "I do n't *gossip* with Harold, Ginny." (JSM1)

(4) She was aware of the impact *that* this declaration made. (AB1)

Some cases are problematic due to "lack of consensus about annotation schemes" (Leech 2011: 168). The highlighted words in examples (5) and (6) are traditionally classified as subordinator and adverb/particle, respectively (e.g. Biber *et al*., 1999: 76). However, they are regarded as prepositions in a number of other frameworks, such as Huddleston and Pullum's (2002: 599, 612). As a consequence of this expansion of the category, prepositions can be complemented by finite clauses and they can be intransitive, i.e. occur without a complement. A similar analysis of Norwegian is found in Holmes and Enger (2018: 322) and in Faarlund *et al*. (1997), which is the basis for the PoS classification in the Oslo-Bergen tagger (Johannessen *et al*., 2012: 57). Allowing for this analysis, examples such as (5) and (6) were retained in the material. Finally, example (7) shows a case of a stranded preposition occurring at a distance from its complement. Such examples were also retained.

(5) Mattie didn't think about the heat *as* she walked beside Butch. (GN1)

(6) … but mostly to re-create that moment when Townsend brought Celia *in*… (AH1)

(7) Han forsøkte å dempe smertene litt ved hjelp av kamferdråper som han masserte både jekselen og tannkjøttet *med*. (EG2)
Lit: "…which he massaged both the molar and the gums with"
He tried to alleviate the pain with camphor drops , which he rubbed into the offending molar and inflamed gum with the tip of his finger. (EG2T)

It may be argued that the PoS-tags in the corpus are to some extent inconsistent – though not necessarily wrong – as those subordinating conjunctions/adverbs that are homonyms of traditional prepositions seem more likely to be tagged as preposition than others – there are for example no instances of *because/fordi* in the samples retrieved. Fortunately, the English and the Norwegian taggers appear to behave relatively similarly in this respect (see Section 5.2). On balance, this potential inconsistency was considered acceptable in view of issues of replicability. The resulting material thus retains the PoS classification assigned by the taggers with only the obvious tagging errors removed (see Table 1 in Section 5.1 for the final size of the dataset).

4. Classificatory framework

Each instance of a 'noun + preposition' sequence was annotated for the syntactic function of the postnominal PP, the complement of the preposition and the general meaning of the PP. The categories are described below.

The classification of syntactic functions is in agreement with categories found in e.g. Biber *et al*. (1999). The following functions were identified: postmodifier (of preceding noun), adverbial (independent of preceding noun), modifier of prenominal adjective, as in (8), part of complex prepositions, e.g. *in front of, ved siden av* ('at the side of' = 'beside'), and part of multi-word verb, as in (9).

(8) … to move to Amsterdam, a *larger city than Leiden*... (JH1)
… å flytte til Amsterdam, en *større by enn Leiden*… (JH1T)

(9) Men alle hadde naturligvis *lagt merke til* det. (EG1)
Lit: "But everyone had naturally *laid mark to* ['noticed'] it."
… no one could have failed to *notice* such outlandish habits … (EG1T)

The meanings of adverbials are identified according to the categories of adjuncts outlined in Hasselgård (2010: 39), while conjuncts and disjuncts are not specified in further detail because of their low numbers. The adjuncts that occur more than once or twice are of the following types:

- **Manner** (including accompaniment and method/means), e.g. …they raised their voices in bright greetings… (BO1); hevet de stemmen i muntre hilsener…

- **Participant**, e.g. Han … hentet varene til henne. (HW1); He… got the groceries for her.

- **Place**, e.g. En gang i uken har hun time på helsesenteret… (BV2); Once a week she has a session at the health centre,…

- **Reason/purpose**, e.g. Sarah would be stopping by the house for the rug. (AT1); Sarah ville komme til huset for å hente teppet.

- **Respect**, e.g. to tell his wife about the journey up the M1. (ST1); …og fortelle kona om turen han hadde fått.

- **Time**, e.g. We were living with my mother for four years, ... (DL2); Vi bodde hos mor i fire år …

The general meanings of postmodifying PPs were identified according to the framework detailed in Hasselgård (2016, 2019), based on Sinclair (1991), although it had to be modified because of the wider scope of the present investigation, i.e. the greater variety of prepositions studied. The meaning categories are an attempt to describe the relationship between the noun preceding the preposition (N1) and the head of the NP complementing the preposition (N2). Those that recur in the material are the following:[7]

- **Argument of nominalization**: The noun is a nominalization and the PP represents an argument (subject, object, adjunct), as in *the presence of death; nærvær av døden* (nom-S), *the lending of money; utlån av penger* (nom-O), *undringen over livet; their astonishment at the world* (nom-A). This category was used whenever the noun was a nominalization, regardless of meaning.

- **Attribute**: The PP specifies a property of the NP head, e.g. *mannen med de store hendene*; *the man with the large hands.*

- **Focus**: The first noun specifies some aspect of the second (Sinclair, 1991: 87), e.g. *et glass med syltetøy*; *a jar of jam.*

- **Locative**: The PP has locative meaning, e.g. *køen ved disken; the queue at the counter*

[7] The examples all come from the material studied. In order to illustrate both languages simultaneously, only examples with congruent translations have been selected here.

- **Part**: a part-whole relationship between N1 and N2, e.g. *the foot of a tree; foten av et tre.*

- **Possessive**: the noun in the PP denotes the possessor of the preceding referent, e.g. *minen til et menneske som…; the expression of someone who…*

- **Quantifier**: The first noun quantifies the second, e.g. *tusener av skritt; thousands of footsteps.*

- **Support**: The noun in the PP carries the most important meaning and is the notional head of the NP, while the preceding noun has a supporting role (Sinclair, 1991: 89), e.g. *various forms of self-advertisment*; *forskjellige former for egenreklame.*

- **Temporal**: the PP has temporal meaning, e.g. *løvtrær om sommeren; green trees in summertime*.

For the study of translations (Section 5.6), the correspondences were classified according to the framework presented in Johansson (2007: 25) as congruent (having the same formal structure as the source), non-congruent (having a different formal structure than the source), and zero (in cases where the 'noun + preposition' sequence had no correspondence in the translation). See Section 5.6 for further explanation and examples.

5. Corpus analysis

This section presents the analysis of the two random samples of 'noun + preposition'. After a general overview of the data, the English and Norwegian prepositions and the types of complementation are surveyed. Then follows an analysis of the syntactic functions of the postnominal PPs before the meanings of postmodifying and adverbial PPs are compared across the two languages. Section 5.6 looks into the translation correspondences of postmodifying and adverbial PPs in both directions of translation (English-Norwegian and Norwegian-English).

5.1 Overview of the data

As detailed in Section 3, random samples of 500 concordance lines were extracted from English and Norwegian original fiction texts. After the exclusion of wrongly tagged hits, the samples were reduced to 474 in English and 475 in Norwegian. For some reason, the random samples come from only 16 out of 30 corpus texts in either language, hence the samples may not be representative of the entire corpus. The overview is laid out in Table 1.

Table 1. Overview of retrieved instances of 'noun + preposition' in ENPC fiction.

	English original	**Norwegian original**
Total number of hits	17,830	17,146
Random sample	500	500
Excluded due to tagging error	26	25
Adjusted sample	474	475
Number of texts (of 30)	16	16

5.2 The structure of PPs in English and Norwegian

The prepositions occurring after nouns are more varied in Norwegian than in English, with 60 types in the Norwegian (adjusted) sample as against 46 in English. As shown in Table 2, the preposition *of* is vastly more frequent than any other preposition in the material, which is due to its "overwhelming pattern of usage being in nominal groups" (Sinclair, 1991: 83). Norwegian does not have any preposition with the same status and versatility in nominal groups (Hasselgård, 2016); hence the frequency of the first item on the Norwegian top 10 list is not very much greater than that of the second, and the frequencies drop much less steeply than in the case of English. Notably, the top item on the Norwegian list is a close correspondence of the second on the English list; in fact, with the exception of *of*, the two lists of prepositions are not very different.

Table 2. The most frequent prepositions in the samples.

	English		**Norwegian**	
1	*of*	207	*i* ('in')	80
2	*in*	58	*på* ('on', 'at')	71
3	*for*	34	*av* ('of', 'off', 'by')	58
4	*with*	28	*med* ('with')	48
5	*on*	23	*til* ('to')	48
6	*from*	16	*for* ('for')	27
7	*at*	14	*fra* ('from')	22
8	*as*	9	*om* ('about', 'if')	11
9	*about*	9	*etter* ('after')	9
10	*by*	9	*over* ('over', 'above')	7
		407		381

The ten most frequent prepositions account for 85.9% of the English sample and 80.2% of the Norwegian sample. Of the remaining preposition types, the following occur more than twice (in order of decreasing frequency): English *into, across, without, before, beside, since, than;* Norwegian *under, ved* ('by'), *der* ('there'), *rundt* ('around'), *hos* ('at'), *mellom* ('between'), *opp* ('up'), *som* ('as'), *foran* ('in front of'), *inne i* ('inside'), *mot* ('against'), *uten* ('without').

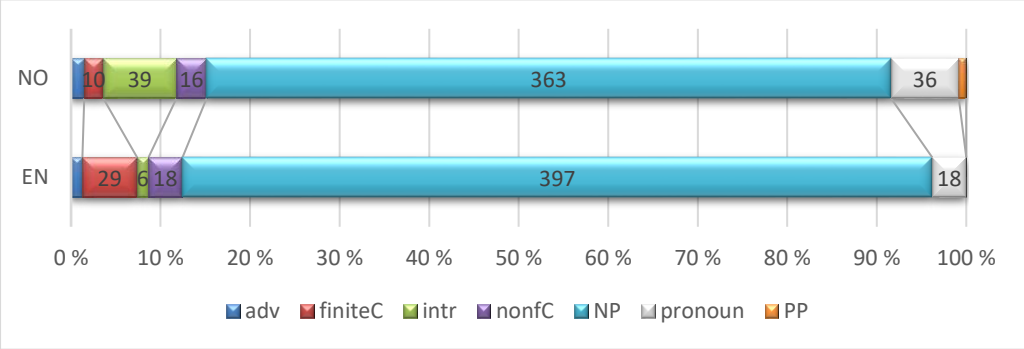

Figure 1. The complementation of the prepositions in Norwegian and English.

Figure 1 shows the distribution of PP complements in the material. Both languages show great preference for noun phrases as complements of the preposition, whether noun-headed (NP) or pronoun-headed (pronoun): 87.5% of the English PPs have a noun phrase as complement and 84% of the Norwegian ones. Intransitive prepositions, as in example (6) above, seem to be more frequent in Norwegian than in English, although this difference is conceivably due to the fact that the languages were tagged with different taggers. Finite clauses are – perhaps surprisingly – more frequent in English, but this is entirely due to instances of the traditional

category of conjunction, as in (10), whose translation replicates the structure of its source. By contrast, some of the Norwegian finite clauses are *at*-clauses (i.e. *that*-clauses), as shown in (11). It should also be noted that the category of non-finite clauses consists exclusively of *å*-infinitives in Norwegian,[8] and mostly of *-ing* participles in English.

(10) I listened to her groans of agony *till we stopped at the edge of a river*… (BO1)
Jeg hørte henne stønne av smerte *til vi stoppet på bredden av en elv*… (BO1T)

(11) Det kunne gå uker og måneder *uten at hun hørte noen nevne det*. (HW1)
Lit: "It could go weeks and months without that she heard anybody mention it"
Weeks, sometimes months, could pass without her hearing anybody say those words. (HW1T)

5.3 The syntactic functions of the PPs

The analysis of the samples produced five different syntactic functions of the postnominal prepositional phrases. These were adverbial, postmodifier of noun, postmodifier of prenominal adjective, part of complex preposition, and part of multiword verb. Those functions that appear in Fang's (2000: 188) list but not in the present study are unlikely to occur directly after a noun, e.g. subject complement. The syntactic functions occur with highly unequal frequencies, as shown in Figure 2, where the functions of adjective modifier, complex preposition and multiword verb have been conflated in the category 'other'.

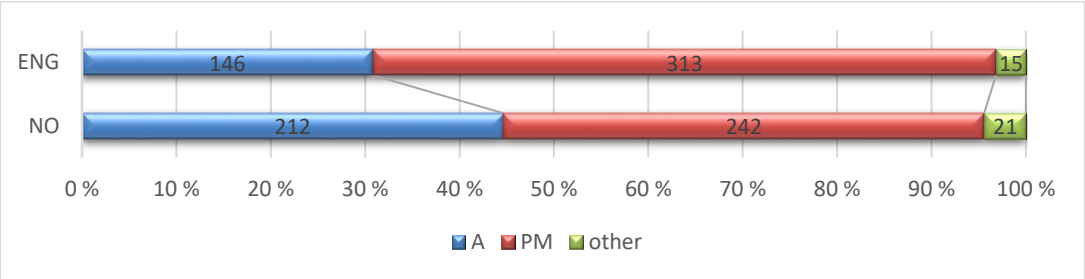

Figure 2. The syntactic functions of postnominal PPs in English and Norwegian (random samples excluding errors).

As the figure shows, the most frequent function of postnominal PPs in both languages is postmodifier (PM). This was expected, as the postnominal position is a favourable context for the postmodifying function. However, the postmodifiers are in even greater majority in English than in Norwegian, where adverbials (A) are rather more frequent. The difference in distribution between the languages is significant at $p<0.001$ (Pearson's chi-squared test: 22.25, DF=2).[9] Because of the very low frequencies of the 'other' categories, the remainder of this paper will focus on postmodifiers and adverbials.

5.4 The meanings expressed by adverbial PPs

Postnominal PPs functioning as adverbials are not part of the same syntagm as the preceding noun, so the meaning categories do not include any relation between the noun and the PP in the sequence. As detailed in Section 4, the classification follows Hasselgård (2010). Only those meanings that occur three times or more in one of the corpora are mentioned separately in

[8] The Norwegian infinitive occurs with or without the infinitive marker *å* in front of the base form of the verb, and is thus structurally similar to the English infinitive (Holmes and Enger, 2018: 226).
[9] The test was carried out using the tools available at http://corpora.lancs.ac.uk/stats/toolbox.php (Brezina, 2018).

Figure 3. The 'other' category includes adjuncts of condition, concession, accompanying circumstance, and comparison. It is important to note that the distribution of adverbial categories shown in Figure 3 applies only in this particular grammatical context, not generally. For the general distribution of adverbials in English fiction, not limited to postnominal position or to prepositional phrases, see Hasselgård (2010: 260−262).

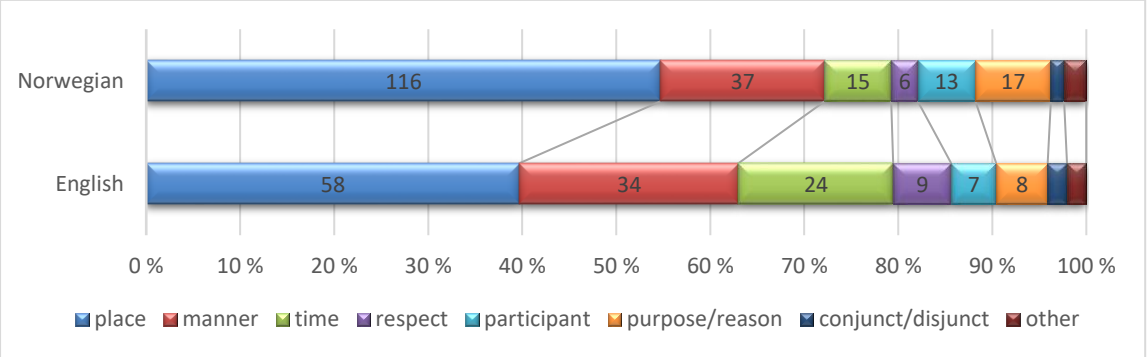

Figure 3. The meanings expressed by postnominal PPs functioning as adverbials in English (N=146) and Norwegian (N=212).

As Figure 3 shows, place adjuncts are the most frequent category in both languages, followed by manner. However, Norwegian has a greater proportion of place adverbials than English does (55% vs. 39%). An example is given in (12).

(12) …det var 95% vann *i en agurk*… (JG1)
Lit: "there was 95% water in a cucumber"
…a cucumber was 95 percent water… (JG1T)

In contrast, English has greater proportions of manner adverbials (24% vs 17%) and time adverbials (16% vs 7%).[10] Examples are given in (13) and (14).

(13) … and stumbled out of the house *in drunken merriment*. (BO1)
…og sjanglet *fulle og lystige* ut. (BO1T)
Lit: "and stumbled drunk and merry out"

(14) We've been living in this motel *for weeks*… (MA1)
Vi har bodd på dette motellet *i mange uker*… (MA1T)
Lit: "We have lived on this motel in many weeks…"

Other categories of adverbials are rather infrequent in both languages, and have more similar proportions.

5.5 The meanings expressed by postmodifying PPs

The meanings expressed by postmodifying PPs are analysed in relation to the preceding noun (see Section 4) and are displayed in Figure 4. While most of the meanings occur with rather similar frequencies in the two languages, a conspicuous difference is the far greater frequency of modifiers with locative meaning in Norwegian, where they account for 37% of the total compared to 18% in English. This is parallel to the situation with adverbial PPs, where spatial adjuncts are more frequent in Norwegian. An example of a locative postmodifier is given in (15), in which the translation mirrors the original.

[10] Time adjuncts are considerably less frequent in postnominal position than would be expected from their general frequency (Hasselgård, 2010: 261), which is due to the colligational restriction on the present dataset: many time adjuncts occur clause-initially or post-verbally (2010: 57).

(15) Hun vinker bak *gardinene i annen etasje*. (BV1)
Lit: "She waves behind the curtains in second floor."
She waves from behind *the curtains on the first floor*. (BV1T)

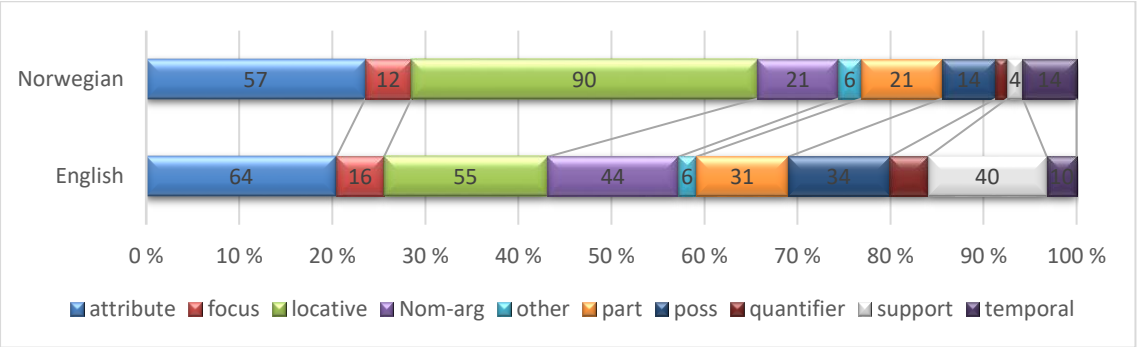

Figure 4. Meanings of postmodifying postnominal PPs in English (N=313) and Norwegian (N=242).

In contrast, English has a greater proportion of postmodifiers with support meaning (13% vs. 2%) and possessive meaning (11% vs. 6%). Examples are given in (16) and (17). While the Norwegian translation has omitted the support noun in (16), the one in (17) follows the English original closely (see further Section 5.6).

(16) Then there was *the matter of her job*. (NG1)
Og så var det arbeidet hennes. (NG1T)
Lit: "And then there was the work hers"

(17) I came to realise that they were *the voices of my spirit companions*. (BO1)
…det var *stemmene til mine følgesvenner* i åndeverdenen. (BO1T)
Lit: "…it was the voices of my companions in the spirit world"

Figure 4 shows a slightly higher percentage of postmodifiers functioning as arguments in NPs with a nominalized head in English (14%) than in Norwegian (8.7%), although the frequency differences are perhaps not as great as might be expected on the basis of Behrens's (2014) comparison of nominalizations in English and Norwegian academic prose. Another possible difference between the languages is the type of arguments that occur as postmodifiers: in English equal proportions of the nominal arguments correspond to objects and adjuncts of clausal constructions, while in Norwegian a larger share correspond to adjuncts; see examples (18) and (19). The proportions of postmodifiers of nominalizations corresponding to subjects, as in (20), are similar. However, the numbers are low, so further study is needed to see if this is a trend.

(18) … she had interrupted *the shedding of her fourth husband* to be present at her son's "first marriage". (AH1) (cp. She shed her fourth husband.)
…hun hadde utsatt å kvitte seg med sin fjerde mann for å være tilstede ved sønnens "første giftermål". (AH1T)
Lit: "she had postponed to get rid of her fourth husband…"

(19) Ingen *flying etter jenter*. (EFH1)
No *running after girls*. (EFH1T) (cp. You must not run after girls.)

(20) The *presence of a man* in the house subdued the women. (ST1) (cp. A man was present…)
Nå da en mann var kommet til stede, dempet kvinnene seg betraktelig. (ST1T)
Lit: "Now that a man had come to the place, the women calmed themselves considerably."

5.6 Translation of postnominal PPs

The aligned translations of the concordance lines were analysed for their degree of congruence with their sources. The categories of correspondence are those outlined in Johansson (2007: 25), where 'congruent correspondence' means that the translated item belongs to the same formal category as that of its source, as in (19) above, while a non-congruent (or 'divergent') translation does not, as in (18) and (20) above. In the present study, congruent translations are those in which the postnominal PP has been translated by a PP with the same syntactic function. Zero correspondence means that the source item has been omitted in the translation. As in the above sections, PPs functioning as adverbials and postmodifiers are presented separately. The results for adverbials are shown in Figure 5. According to a Pearson's Chi-squared test, the two directions of translation differ only marginally: $\chi^2 = 5.84$ (2), p = 0.05.

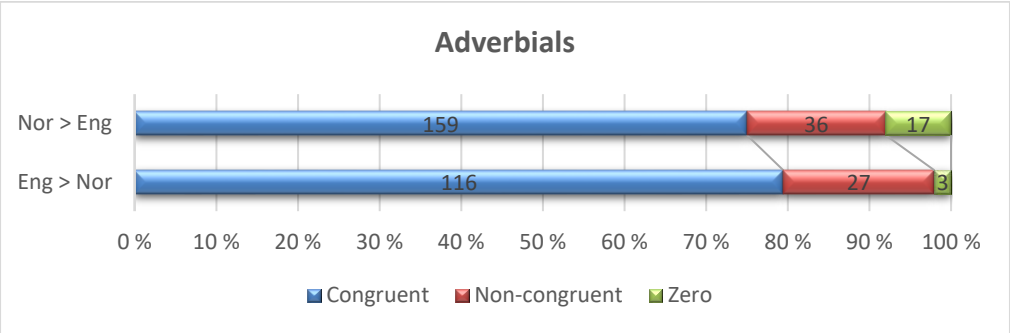

Figure 5. The translation of adverbial postnominal PPs.

The percentage of congruent translations of adverbial PPs is extremely high in both directions of translation: 74.5% in translations from Norwegian to English and 79.5% in translations from English to Norwegian. This indicates that the resources as well as the preferences for forming adverbials in postnominal position are similar in the two languages. Although the difference between the two directions of translation is only borderline significant, the percentage of zero correspondence is noticeably higher in translations from Norwegian. The zero cases include some idiomatic expressions, such as *se for seg* ('see before one', 'imagine') as in example (21). In a few cases, the translation is so free that there is little trace of the syntax of the original, and sometimes the PP is omitted for no apparent reason, as in (22).

(21) Tora pleide å se *Gunn for seg* når hun lå alene i kammerset sitt om kvelden og ikke fikk sove. (HW1)
Lit: "Tora used to see Gunn before her when she lay alone…"
At night, when she was alone in her room and couldn't sleep, Tora would sometimes *see Gunn*. (HW1T)

(22) For ungdommen skal reise over sundet *med ferja* for å sjå på film… (EH1)
Lit: "For the young shall travel over the sound with the ferry for to see film"
The young people are crossing the sound [Ø] to go to the movies… (EH1T)

Non-congruent correspondences represent a variety of constructions, including differences in syntactic realization, as in (23) and (24), and in lexicalization, as in (25).

(23) Hvis vi retter det samme spørsmålet *til en som fryser*, er svaret varme. (JG1)
Lit: "If we direct the same question *at one who freezes*, is the answer warmth."
If we ask *someone dying of cold*, the answer is warmth. (JG1T)

(24) Aristotle remembered that such busts of Homer were common in Thessaly, Thrace, Macedonia, Attica, and Euboca *in his lifetime*. (JH1)

Aristoteles husket at slike byster av Homer var alminnelige i Tessalia, Trakia, Makedonia, Attika og Euboia *da han levde*. (JH1T)
Lit: "… when he lived"

(25) Fikk fru Olsrud noe brev *i det siste*? (EG1)
Lit: "Got Mrs Olsrud any letters *in the last*?"
Had she had any letters *lately*? (EG1T)

While the correspondences of adverbial PPs show a high degree of similarity between the languages and between the directions of translation, the analysis of postmodifying PPs indicate greater cross-linguistic differences. The degree of congruence is shown in Figure 6. According to a Pearson's Chi-squared test, the two directions of translation differ significantly: $\chi^2 = 23.66$ (2), $p < 0.00001$.

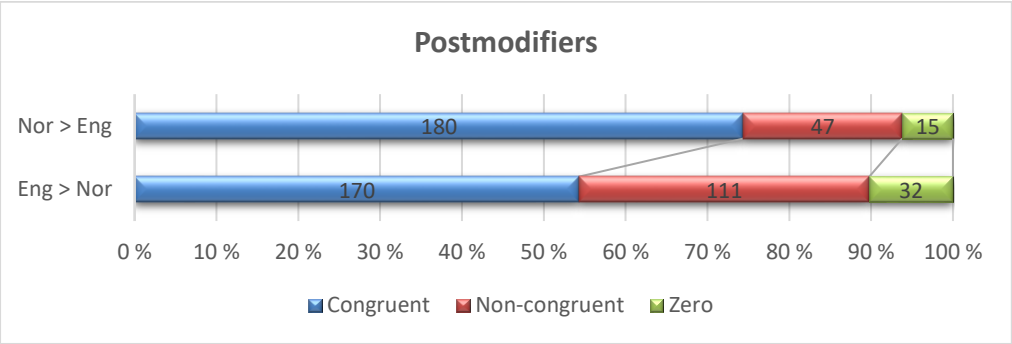

Figure 6. The translation of postmodifying postnominal PPs.

A striking feature of Figure 6 is the much higher proportion of congruent correspondences in translations from Norwegian into English than in the opposite direction. Hence, it appears to be easier to transfer the Norwegian pattern of postmodification into English than to translate English postmodifiers into Norwegian. The proportion of congruent English translations of Norwegian postmodifiers is 74.4%, which is practically the same as in the case of congruent adverbials. Of English postmodifiers translated into Norwegian, by contrast, only 54.3% are congruent. Close examination of the English postmodifiers with non-congruent Norwegian correspondences reveals that the culprit is *of*: 71% of the non-congruent translations (79 out of 111) have *of* in the source. *Of* is also responsible for 81% of the zero correspondences (26 out of 32), but only 48% (82 out of 170) of the congruent correspondences.[11] There is no similarly discernible feature that can explain non-congruence in translation from Norwegian into English.

Prepositional phrases with *of* function predominantly as postmodifiers of nouns; only nine out of the 207 cases have other functions (respect adjunct, part of complex preposition, and part of multi-word verb). The most frequently occurring Norwegian preposition in congruent translations of *of* is its cognate *av* (46 occurrences, 22%). Conversely, 37 out of 46 instances of *av* in postmodifying PPs (80%) are translated by *of*. The most frequent meaning expressed in the congruent correspondences between *av* and *of* is the partitive one, shown in (26).

(26) Dermed smeller dørene igjen i dypet *av skipet*. (EFH1)
Lit: "Then slam the doors shut in the depth of the ship."
The doors slam shut in the bowels *of the ship*. (EFH1T)

[11] This is in line with the findings of Hasselgård (2014: 64), where '*the* N1 *of the* N2' had congruent correspondences in only 43.5% of Norwegian translations and 33% of Norwegian sources.

Another typical meaning of *of*-phrases is possession, i.e. the *of*-genitive. Norwegian has a similar periphrastic genitive using the preposition *til* (as an alternative to the *s*-genitive, like in English). However, the postmodifiers with possessive meaning tend to occur among divergent modifiers in both directions of translation: the prepositional genitives are translated congruently in less than a third of the cases. Non-congruent examples are given in (27) and (28).

(27) Broren *til Tora* sykler med blomsterpakker… (BV2)
Lit: "The brother *of Tora* cycles with flower-parcels"
Tora's brother delivers flowers by bicycle. (BV2T)

(28) David and Harriet were commended for their fertility, and jokes were made about the influences *of their bedroom*. (DL1)
David og Harriet ble rost for sin fruktbarhet, og man spøkte om *soverommets innvirkning*. (DL1T)
Lit: "…and one joked about the bedroom's influence."

The factors governing the choice between the *s*-genitive and the periphrastic genitive seem to differ between the languages. In Norwegian, the choice is to a large extent governed by formality (Holmes and Enger, 2018: 49). In both languages, the *s*-genitive is considered the more formal alternative, which is less likely to occur in speech. In English, an important additional factor is whether the possessor is human or non-human, although register also plays a role (Biber *et al*., 1999: 302). The periphrastic genitive is used with a human possessor in the Norwegian original of (26), while the English translator prefers the *s*-genitive. In (27), the possessor is non-human, but the style is rather formal, so English has the *of*-genitive and Norwegian the *s*-genitive. It may be noted that not all divergent correspondences of possessives are *s*-genitives; they may also be, for example, locative modifiers and relative clauses.

6. Discussion and concluding remarks

In Section 1, the following research questions were asked:

- What are the syntactic functions of PPs following a noun in Norwegian and English?
- What meanings do the PPs convey?
- Are there quantitative and qualitative differences between the languages as regards the functions and meanings of postnominal PPs?
- To what extent are translations congruent?

The analysis has shown that the same syntactic functions are found in the two languages, but with different frequencies. The 'noun + preposition' sequence represented postmodification of nouns and prenominal adjectives, clause-level adverbials, and components of complex prepositions and multi-word verbs. The greatest cross-linguistic difference concerns the overall proportions of adverbials and postmodifying PPs, where – as expected – English had a larger proportion of postmodifiers while Norwegian had a larger proportion of clause-level adverbials. The functions of PPs following nouns may thus support the claim (Nordrum, 2007; Behrens, 2014) that English is more nominal and Norwegian is more clausal. It may be noted that the higher frequencies of English postmodifiers resonate with the findings of Moreira-Rodríguez (2006) and Mott (2013) who both conclude that the postmodifying function of PPs is more flexible in English than in Spanish.

In terms of the meanings expressed, Norwegian postnominal PPs were locative more frequently than English ones in both adverbial and postmodifiying functions. Among the postmodifiers, English had more PPs modifying support nouns (Sinclair, 1991) as well as possessive constructions and modifiers of nominalizations. English also had larger proportions of adverbials with temporal and manner meanings. Other meanings were more similarly distributed.

The degree of congruence in translation differs across translation directions and syntactic functions. Adverbials are congruent in 75-80% of the cases in both directions of translation. Postmodifiers are congruent more often in translation from Norwegian to English (c. 74%) than from English to Norwegian (c. 54%). This indicates greater cross-linguistic differences in postmodifying than in adverbial PPs, and as a consequence, that postmodifying PPs are less straightforward for translators than adverbials are. Closer analysis showed that much of the difference regarding postmodifying PPs can be attributed to the special position of the highly frequent preposition *of* in combining "with preceding nouns to produce elaborations of the nominal group" (Sinclair, 1991: 83) and the great variety of meanings and relations that it can express. The fact that *of* lacks an equivalent in Norwegian can make Norwegian translators more inclined to change the structure. On the other hand, the translation of Norwegian postmodifiers into English does not involve any similarly difficult structure, which appears to make English translators more likely to keep the source structure.

Some of the findings of the present study must be regarded as tentative due to the limited scope of the investigation. The question of whether English is more nominal than Norwegian still remains to be resolved, even if the present study points in the same direction as e.g. Behrens (2014) and Nordrum (2007). However, future investigations need to cover more material and more registers in both languages. The finding that Norwegian PPs express locative meanings more often than English ones in both adverbial and postmodifying functions is interesting and should be followed up by studying PPs more generally without the restriction of a preceding noun, possibly also in comparison with other types of locative expression. The apparent language contrast in the conditions governing the periphrastic genitive vs. the *s*-genitive is another thread worth pursuing. Yet another is the detection of patterns in the sense of Hunston and Francis (2000) among those 'noun + PP' sequences that represent postmodification or complementation of nouns.

Finally, the chosen method of using a PoS-tag sequence as the basis of a cross-linguistic investigation proved to be productive in identifying cross-linguistic similarities and differences. As discussed in Section 3, the fact that different taggers were used for the two languages is a potential problem. In the present study, this problem was minimized by the fact that the structures that involved conflicting or controversial tags were relatively infrequent and hence will have had little impact on the main findings. It would, of course, have been reassuring for comparability if the two taggers had used the same tag set and the same PoS definitions, which might have been possible when the two languages are as closely related as Norwegian and English are (though see some misgivings voiced by Johansson (2007: 306)). However, the use of tag sequences rather than lexically defined searches also constitutes a kind of bottom-up procedure with the potential to retrieve instances and uses which might not have surfaced otherwise. Handled with care, tag sequences may indeed act as a window into cross-linguistic similarities and differences in lexicogrammar.

Corpus

English-Norwegian Parallel Corpus (ENPC), fiction. http://www.hf.uio.no/ilos/english/services/omc/, Accessed through Glossa at https://tekstlab.uio.no/glossa2/omc4 [Last accessed 2 June 2021].

References

Behrens, B. 2014. A Case Study of Linguistic Text Conventions in Comparable and Parallel texts: English and Norwegian. In *Corpus-based Studies in Contrastive Linguistics*, *Oslo Studies in Language* 6(1), S.O. Ebeling, A. Grønn, K.R. Hauge and D. Santos (eds), 143–160. http://www.journals.uio.no/osla [Last accessed 2 June 2021].

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

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Egan, T. 2013. *Between* and *through* Revisited. In *Corpus Linguistics and Variation in English: Focus on Non-native Englishes*, M. Huber and J. Mukherjee (eds). *VARIENG: Studies in Variation, Contacts and Change in English*, Vol. 13. http://www.helsinki.fi/varieng/journal/volumes/13/egan/ [Last accessed 2 June 2021].

Elsness, J. 2014. Clausal Modifiers in Noun Phrases: A Comparison of English and Norwegian Based on the Oslo Multilingual Corpus. In *Corpus-based Studies in Contrastive Linguistics*, *Oslo Studies in Language* 6(1), S.O. Ebeling, A. Grønn, K.R. Hauge and D. Santos (eds), 91–118. http://www.journals.uio.no/osla [Last accessed 2 June 2021].

Faarlund, J.T., Vannebo, K.I. and Lie, S. 1997. *Norsk Referansegrammatikk*. Oslo: Universitetsforlaget.

Fang, A.C. 2000. A Lexicalist Approach towards the Automatic Determination for the Syntactic Functions of Prepositional phrases. *Natural Language Engineering* 6(2): 183–201.

Granger, S. and Bestgen, Y. 2014. The Use of Collocations by Intermediate vs.Advanced Non-native Writers: A Bigram-based Study. *International Review of Applied Linguistics in Language Teaching*, Volume 52(3): 229–252.

Granger, S. and Rayson, P. 1998. Automatic Profiling of Learner Texts. In *Learner English on Computer,* S. Granger (ed.), 119–131. London: Longman.

Hasselgård, H. 2010. *Adjunct Adverbials in English*. Cambridge: Cambridge University Press.

Hasselgård, H. 2014. Discourse-structuring Functions of Initial Adverbials in English and Norwegian News and Fiction. *Languages in Contrast* 14(1): 73–92 . doi: 10.1075/lic.14.1.05has

Hasselgård, H. 2016. *The way of the world*: The Colligational Framework 'the N1 of the N2' and its Norwegian Correspondences. *Nordic Journal of English Studies* 15(3): 55–79.

Hasselgård, H. 2019. *The nature of the essays*: The Colligational Framework 'the N of the N' in L1 and L2 Novice Academic English. In *Corpus Approaches into World Englishes and Language Contrasts*, H. Parviainen, M. Kaunisto and P. Pahta (eds). *Studies in Variation, Contacts and Change in English*, Vol. 20. https://varieng.helsinki.fi/series/ [Last accessed 2 June 2021].

Hasselgård, H. 2021. Time Adverbials in English and Norwegian News Discourse. In *Time Languages, Languages in Time*, A. Čermáková, T. Egan, H. Hasselgård and S. Rørvik (eds), 201–227. Amsterdam/Philadelphia: John Benjamins.

Holmes, P. and Enger, H.-O. 2018. *Norwegian. A Comprehensive Grammar*. London / New York: Routledge.

Huddleston, R. and Pullum, G.K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. 2000. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.4

Johansson, S. 2007. *Seeing through Multilingual Corpora*. *On the Use of Corpora in Contrastive Studies*. Amsterdam / Philadelphia: John Benjamins. https://doi.org/10.1075/scl.26

Johannessen, J.B., Nygaard, L., Priestley, J. and Nøklestad, A. 2008. Glossa: a Multilingual, Multimodal, Configurable User Interface. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. 2008, 617–622. Available at http://urn.nb.no/URN:NBN:no-46163 [Last accessed June 2020].

Johannessen, J.B., Hagen, K., Lynum, A. and Nøklestad, A. 2012. OBT+stat. A Combined Rule-based and Statistical Tagger. In *Exploring Newspaper Language. Corpus compilation and research based on the Norwegian Newspaper Corpus*, G. Andersen (ed.), 51–65. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.49.03joh

Leech, G.. 2011. Principles and Applications of Corpus Linguistics. In *Perspectives on Corpus Linguistics*, V. Viana, S. Zyngier and G. Barnbrook (eds), 155–170. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.48.10lee

Moreira-Rodríguez, A. 2006. 'The book *on the table*', 'The man *in the moon*': Post-modification of Nouns by Preposition + Noun in English and Castilian. *Bulletin of Spanish Studies* 83(1): 53−72. DOI: 10.1080/1475382062000346045

Mott, B. 2013. Postmodifying Prepositional Phrases in English and Spanish (with Special Reference to Locative Postmodifiers). *Transfer* 8(1–2): 153–170. DOI: https://doi.org/10.1344/transfer.2013.8.153-160.

Nordrum, L. 2007. *English Lexical Nominalizations in a Norwegian-Swedish Contrastive Perspective*. PhD thesis, University of Göteborg. https://gupea.ub.gu.se/bitstream/2077/17181/5/gupea_2077_17181_5.pdf [Last accessed June 2020].

Renouf, A. and Sinclair, J. 1991. Collocational Frameworks in English. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, K. Aijmer and B. Altenberg (eds), 128–144. London: Longman.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Wilhelmsen, A. 2019. Pretty complete *or* completely pretty*? Investigating Degree Modifiers in English and Norwegian Original and Translated Text*. MA thesis. Faculté de philosophie, arts et lettres, Université catholique de Louvain, 2019. http://hdl.handle.net/2078.1/thesis:18894 [Last accessed June 2020].

Author's address

Hilde Hasselgård
Department of Literature, Area Studies and European Languages
University of Oslo
P.O. box 1003, Blindern
NO-0315 Oslo
Norway
hilde.hasselgard@ilos.uio.no

RAF, *DNA* and *CAPTCHA*: English acronyms in German and Swedish translation

Jenny Ström Herold, Magnus Levin, Jukka Tyrkkö

Linnaeus University (Sweden)

This study investigates acronyms in English originals and their translations into German and Swedish, comparing forms, functions and distributions across the languages. The material was collected from *the Linnaeus English-German-Swedish corpus* (*LEGS*) consisting of original and translated popular non-fiction. From a structural point of view, acronyms most often occur as independent noun heads (*When IBM introduced* […]) or as premodifiers in a noun phrase (*PGP encryption*). Due to morphosyntactic differences, English acronym premodifiers often merge into hyphenated compounds in German translations (*UN-Klimakonvention*), but less frequently so in Swedish. The study also discusses explicitation practices when introducing source-culture specific acronyms in the translations. German translators explain and elaborate more than Swedish translators and they do so in the German language. Swedish translators, however, use English to a greater extent, suggesting that Swedish readers are expected to have better knowledge of English than German readers.

Keywords: acronyms, abbreviations, translation, explanation practices, explicitation, LEGS, compounds, premodifiers, English/German/Swedish

1. Introduction

Acronyms are prevalent and ever more frequent in English (Xu *et al*., 2007; Leech *et al*., 2009: 212), German (Kobler-Trill, 1994) and Swedish (Sigurd, 1979: 7; Nübling and Duke, 2007: 231), a development mirroring the increasing societal prominence of science/technology and politics/business outside specialised domains (Kobler-Trill, 1994: 200). For translators, however, acronyms may pose a challenge, especially when they are strongly tied to the source-language culture (Ingo, 2007: 121–122). In spite of this, very little research has been carried out on acronyms from a translation perspective.

Examples (1)–(3) illustrate some of the variation in the translation strategies for acronyms in the data from the Linnaeus University English-German-Swedish corpus (LEGS). In the text where (1) occurs, both the English original and the Swedish translation consistently use the acronym, while the German translation sometimes uses the acronym and sometimes, as in (1b), opts for the spelt-out, explicit form.

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.v11i1.3443.

(1) a. The *RAF* began flying over Germany, […] (LEGS; EN original)
b. *Die Royal Air Force* nahm Flüge über Deutschland auf […] (GE translation)
"The Royal Air Force took up flights over Germany"
c. *RAF* började fälla flygblad […] (SW translation)
"RAF began dropping leaflets"

Acronyms may also be well known in both the source and the target cultures, and such examples are unlikely to cause problems for translators. Some internationally established acronyms may even be more recognisable than their spelt-out forms (Nuopponen and Pilke, 2016 [2010]: 63), as *DNA* in (2).

(2) a. *DNA* tests (EN original)
b. *DNA*-Tests (GE translation)
c. *DNA*-tester (SW translation)

Other instances, however, are more complex and less straight-forward. In (3), the English original itself includes a spelt-out variant of the acronym in brackets. The German translation in (3b) is highly explicit, keeping the English explanation and also adding a German version. The Swedish translation in (3c) instead resorts to a rephrased Swedish version of the original explanation.

(3) a. Complete the CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), […] (EN original)
b. Dann muss ich nur noch den CAPTCHA durchlaufen (den „Completely Automated Public Turing test to tell Computers and Humans Apart", also den „vollautomatischen öffentlichen Turingtest zur Unterscheidung von Computern und Menschen") […] (GE translation)
"i.e. the "completely-automated public Turing-test for distinction between humans and computers"'
c. Fyll i captcha-rutan (ett robotfilter för att skilja människor från datorer), […] (SW translation)
"fill in the captcha-box (a robot-filter to distinguish humans from computers)"

As illustrated in (1)–(3) above, English acronyms occur in different syntactic contexts and as such may function as noun phrase heads and as premodifiers.[1] In German and Swedish, acronyms may also be used independently as heads (as in (1c)) or – a typical solution – integrated into compound nouns as in (2b) and (2c). Another important feature of acronyms concerns their reference, involving different semantic categories. They may thus refer to, for instance, organisations, as in (1), or denote medical or technical terms, as in (2) and (3).

In view of the observed grammatical and semantic flexibility of acronyms in originals and translations and the different options facing translators, this paper investigates both acronym use in English original non-fiction and preferences concerning the translation strategies in German and Swedish target texts. More specifically, it will address the following questions:

- What semantic categories and syntactic functions of acronyms occur in English popular non-fiction and how do these relate to German and Swedish translation correspondences?

[1] Needless to say, the status of 'compounds' or noun sequences in English has been the subject of much discussion (e.g., Giegerich, 2004). In this paper we treat a structure such as *DNA tests* as consisting of a head noun and a noun premodifier, i.e. a noun sequence.

- How are English acronyms introduced and explained in German and Swedish translations?

- What effect, if any, do semantic categories and type frequency differences have on the choice of translation correspondences?

In the following, the term 'acronym' covers both short forms read out as words, or 'true acronyms' (e.g., *NATO* from *North Atlantic Treaty Organization*), and 'initialisms', which are read out letter by letter (e.g., *UK* from the *United Kingdom*) (see Gale, 2007).[2]

The paper is structured as follows. Section 2 gives a brief overview of previous translation-oriented observations on acronyms. This is followed in section 3 by a description of the trilingual corpus used, and the data retrieval methods. Section 4 presents the results, regarding both source-text and target-text usage.

2. Background

The question how acronyms can or should be translated is rarely addressed in translation studies. Ingo (2007: 121–122), however, acknowledges that acronyms can be challenging for translators for a number of reasons. First of all, the translator must pay attention to target-language conventions as when the target-language acronym (*UN* for *United Nations*) is different from the source-language acronym (cf. *FN* for *Förenta Nationerna* in Swedish) or the source-language acronym (Ge. *BRD*) lacks a corresponding acronym in the target language (Sw. *Västtyskland* [West Germany]). In addition, Betancourt Ynfiesta, Treto Suárez and Fernández Peraza (2015: 95) point out that the existence of more than one referent for an acronym may cause difficulties. An example is *AA*, for which the *Oxford English Dictionary* lists five different meanings: *administrative assistant*, *Alcoholics Anonymous*, *anti-aircraft*, *Associate of Arts* and *Automobile Association*.[3] This acronym underlines Ingo's (2007: 121) point that "what you gain in brevity and space, you lose in clarity" [our translation]. Ingo (2007: 123) makes an additional remark which clearly suggests the need for more in-depth studies. When encountering culture-specific acronyms, such as acronyms referring to political parties, the translator has to make additions in the translation to make it understandable for the target reader. However, Ingo does not elaborate further on this.

From a syntactic-morphological point of view, prior observations on contrastive differences are again limited in nature. For instance, Magnusson (1987: 91) suggests that *US-* in German compounds (*der US-Botschafter* ['the US-ambassador']), common in German journalese, should preferably be translated into a Swedish adjective (*den amerikanska ambassadören* ['the American ambassador']). A more extensive corpus study by Ström Herold and Levin (2019: 842) indicates that acronyms are frequently used as premodifiers in English (*WTO ruling*) and are also common as left-hand elements in German compounds (cf. also Fleischer and Barz, 2012: 283), but less so in Swedish. Their frequent use as premodifiers in English can be attributed to their syntactic flexibility. In contrast to the spelt-out form (**Organization for Security and Cooperation in Europe monitors*), the one-word format readily allows premodification (*OSCE monitors*) (cf. Fleischer 1997: 189).

[2] Apart from the typical true acronyms and initialisms, there are some rare hybrid forms which are partly read as words and partly as individual letters, such as *PNAC* (/ˈpɪːnæk/; the *Project for a New American Century*).
[3] A further example is the acronym *CAR*, for which Ehrmann *et al.* (2013: 238) identify ten different referents in their news corpus.

The observations presented above indicate the fragmentary state of current knowledge. Nevertheless, they will serve as useful starting points for our corpus study on English acronyms in translation. Section 3 describes the material and methods used.

3. Material and method

The primary data, comprising 1,699 acronyms from English source texts and their German and Swedish translation correspondences, was collected from the LEGS corpus (Ström Herold and Levin, 2018; 2019), a trilingual translation corpus consisting of popular non-fiction books written in one of the languages and translated into the other two. Genres covered include popular science, biography and history books. This study is based on ten English original texts sampled from the beginning of each book. Each author and translator is represented only once each to avoid any translator or author biases. The English originals were all published in the 2010s and comprise 543,000 words. A main advantage of LEGS is that it allows the comparison of two target languages, which means that target-language-specific preferences can be studied.

The choice of material was guided by both availability and suitability for the given research questions. The most technical genres such as hard-core natural sciences, where one would also expect a high acronym density (cf. Mair, 2006: 62), are generally not translated from English to other languages. The more popularised LEGS genres are those being widely translated today and, as seen in the present study, acronyms are a quite prevalent here as well. A key difference between hard-core and popularised genres is that the latter addresses a broader audience, which means that translators need to consider factors relating to the target readers' degree of knowledge. Thus, the translation strategies for acronyms will most likely reflect not only structural preferences between the target languages but also pragmatically motivated differences relating to target-culture adaptations.

The acronyms were retrieved from the corpus using a script written in Python. When operationalising the retrieval algorithm, we took care to be inclusive of rare occurrences with lower-case letters such as *fMRI* (*functional Magnetic Resonance Imaging*) and with numbers such as *BRCA1* (*Breast Cancer 1*) by defining acronyms as items with at least two consecutive capital letters, which may contain one or more full stops (e.g., *U.S.A.*). The forms with and without full stops were treated as one type, e.g. *USA* and *U.S.A.* We did not include abbreviations such as *APR* (*April*) and *DR* (*Doctor*) on the grounds that they are shortened forms of words and not acronyms in the true sense. Altogether 212 unique acronyms were identified in the primary data.

To examine possible effects of acronym frequencies on explanation practices in translations, we obtained the occurrences of these acronyms in contemporary English, using their relative frequencies in Google Books (UK).[4] A Livecode script was written to run an API call to the Google Ngram Viewer for each acronym in the date range 1990 to 2000. The mean frequency of each acronym during this ten-year period was calculated in order to establish how common the acronym was in written British English. The frequencies were divided into three frequency bands that were used to determine the extent to which the translators' likelihood of explaining acronyms could be accounted for by the frequencies of the acronyms they encountered.

[4] Although the composition of Google Books is sometimes criticised for bias in favour of non-fiction writing (see Pechenick *et al.*, 2015), this does not complicate the comparison in the present case as the LEGS corpus itself comprises exclusively non-fiction texts.

4. Results

Section 4.1 begins with an overview of the distributions of semantic categories identified in the English originals. 4.2 discusses the different syntactic functions in originals, 4.3 focuses on the distributions of translation correspondences in translations, and, finally, 4.4 analyses explanations and language choice in translations.

4.1 Semantic categories and their distributions in English originals

In the material, we identified five major semantic categories from the 1,699 English acronyms (31/10,000 words): 1) measure, 2) medical, 3) organisation, 4) place, 5) technical, and 6) other. Table 1 provides an overview of these categories with examples from LEGS.

Table 1. Semantic acronym categories identified in LEGS.

Category	Examples
measure	*BCE*; *IQ*
medical	*ADHD*; *DNA*
organisation	*ANZAC*; *IBM*
place	*UK*; *US*
technical	*GPS*; *WMD*
other	*CEO*; *OMFG*

The 'measure' category comprises types that potentially occur as units with numbers (e.g., *c. 1700 BCE*). 'Medical' and 'technical' acronyms refer to terminology within these two specialised domains, such as the names of diseases or technical devices. The 'place' category comprises few types, some of which are highly frequent, that refer to toponyms as exemplified in the table. The 'organisation' category includes the names of companies and various national and international organisations. Culture-specific acronyms are mostly found in the final category and, as will become evident below, these pose the main challenge for translators because they often lead to different kinds of adaptions in translations, such as using a cultural equivalent, a functional equivalent (i.e., a generalising paraphrase) or using notes or glosses (see Newmark, 1988: 82–83; 92). The miscellaneous category 'other' comprises mainly business terms and internet slang.

Figure 1 shows the token frequencies of the semantic categories exemplified in Table 1. As also found by Leech *et al.* (2009: 212), the largest category of acronyms involves names of organisations. Place names, which were disregarded by Leech *et al.*, form the second largest group in terms of tokens, while the remaining categories are rarer.

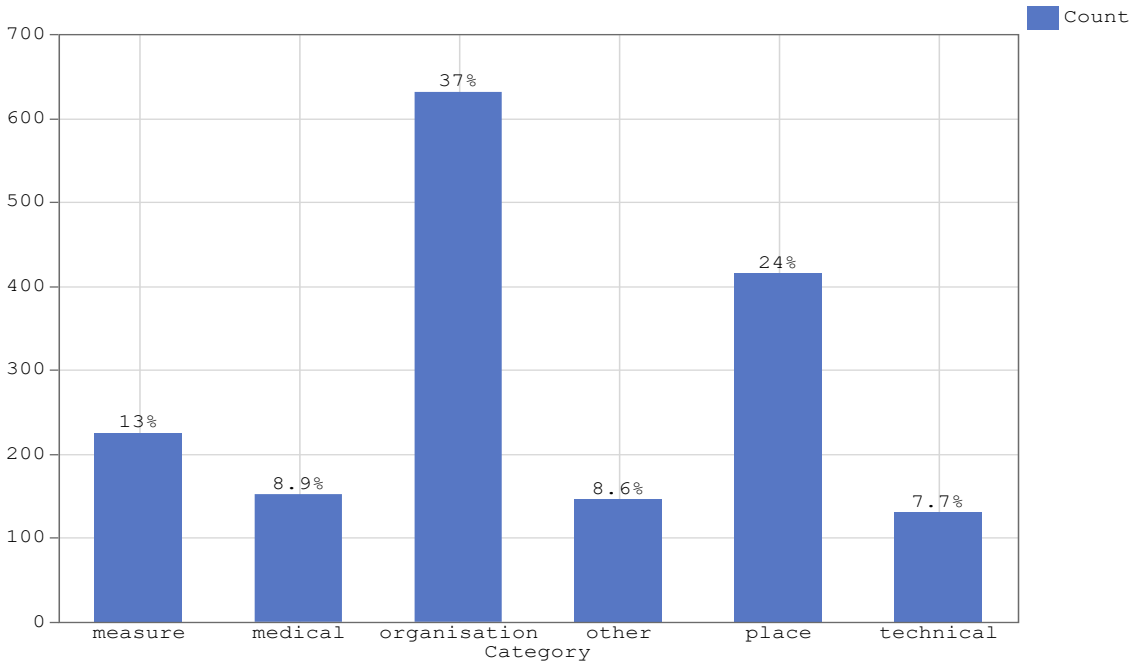

Figure 1. Distribution of semantic acronym categories in LEGS.

The individual acronym type distributions produce a partly different picture, as illustrated in Figure 2 below. To begin with, organisations not only represent the largest number of tokens, but also comprise by far the largest number of types with 107 unique types out of the 212 in the whole dataset. The technical (31 types) and medical (n=19) categories are also reasonably numerous, while place names (n=4)[5] and measures (n=5)[6] comprise very few types but are rather frequent in token counts.

[5] The four types are *UK*, *US*, *USA* and *(Washington) DC*.
[6] The five types are *BCE*, *CE*, *GDP*, *IQ* and *BP* (Before Present).

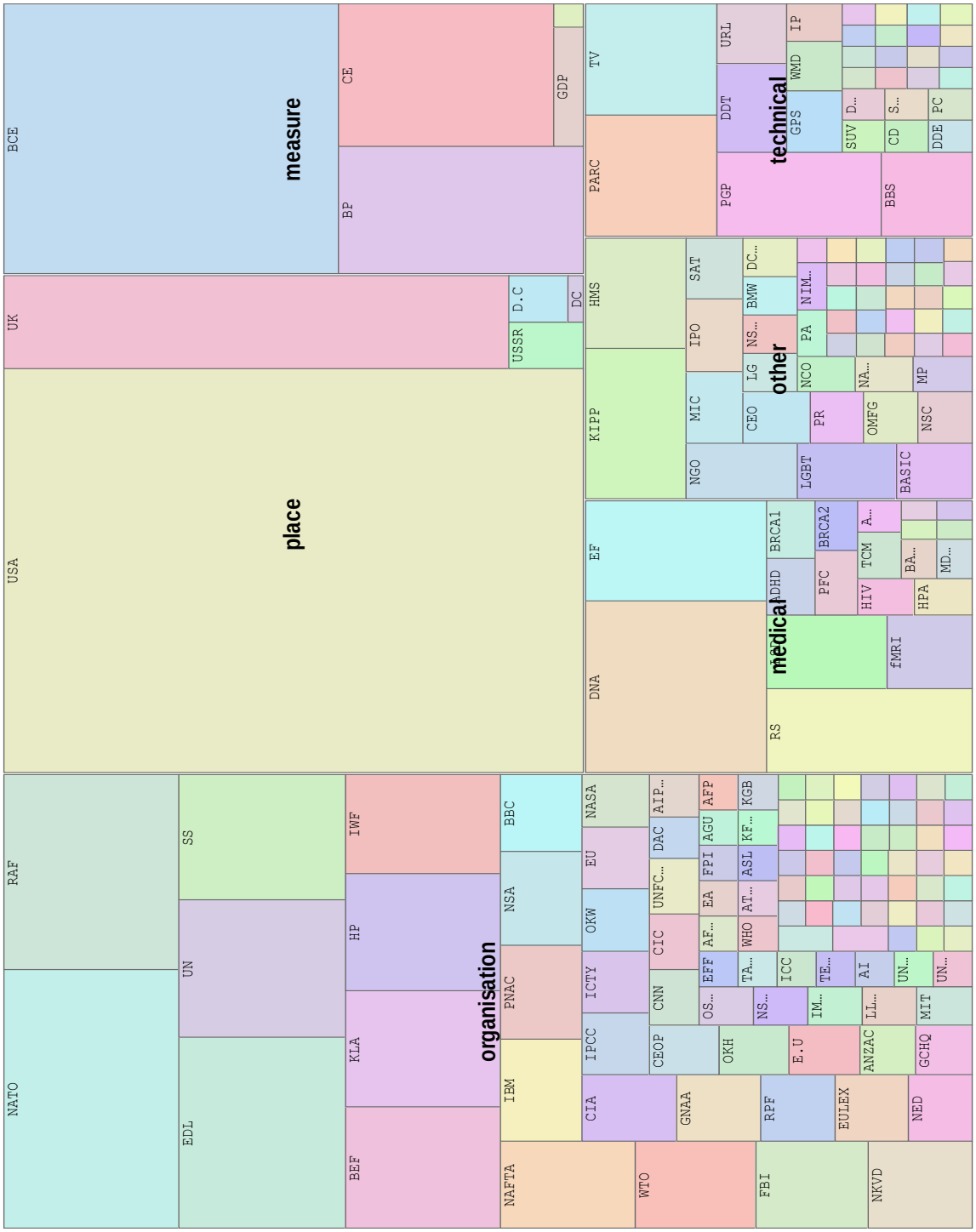

Figure 2. Relative frequencies of individual acronym types by semantic category in LEGS.

The LEGS data thus show that there are considerable frequency differences across semantic categories and acronym types. The two largest semantic categories, organisations and places, differ greatly in their type distributions, and, as will be seen in the next section, also in their syntactic functions.

4.2 Syntactic functions of acronyms in English originals

In the English originals, acronyms fulfil two major and three minor syntactic functions, the two most frequent being noun phrase heads and premodifiers, and the three rarer being

postmodifiers, genitives and compounds. The two major functions, noun-phrase heads and premodifiers, are exemplified in (4) and (5) below:

(4) the military-industrial complex (*MIC*) (EN original)

(5) *EDL* supporters (EN original)

The three minor functions are rare or restricted in use. What we have termed 'postmodifiers' can be seen in (6). Most of these involve two specific time-denoting acronyms: *CE* (*Common Era*) and *BCE* (*Before Common Era*). Even rarer are genitives[7] (as in 7) and compounds (as in 8), in which the acronyms typically are hyphenated with *ed*-participles.

(6) the third century *CE* (EN original)

(7) *CIC's* vision (EN original)

(8) The *U.K.*-based Tax Justice Network (EN original)

Figure 3 presents the syntactic functions of acronyms in correlation with semantic categories. Noun-phrase head is the most common function in the corpus, but, as seen in the mosaic plot below, there are differences between the semantic categories.

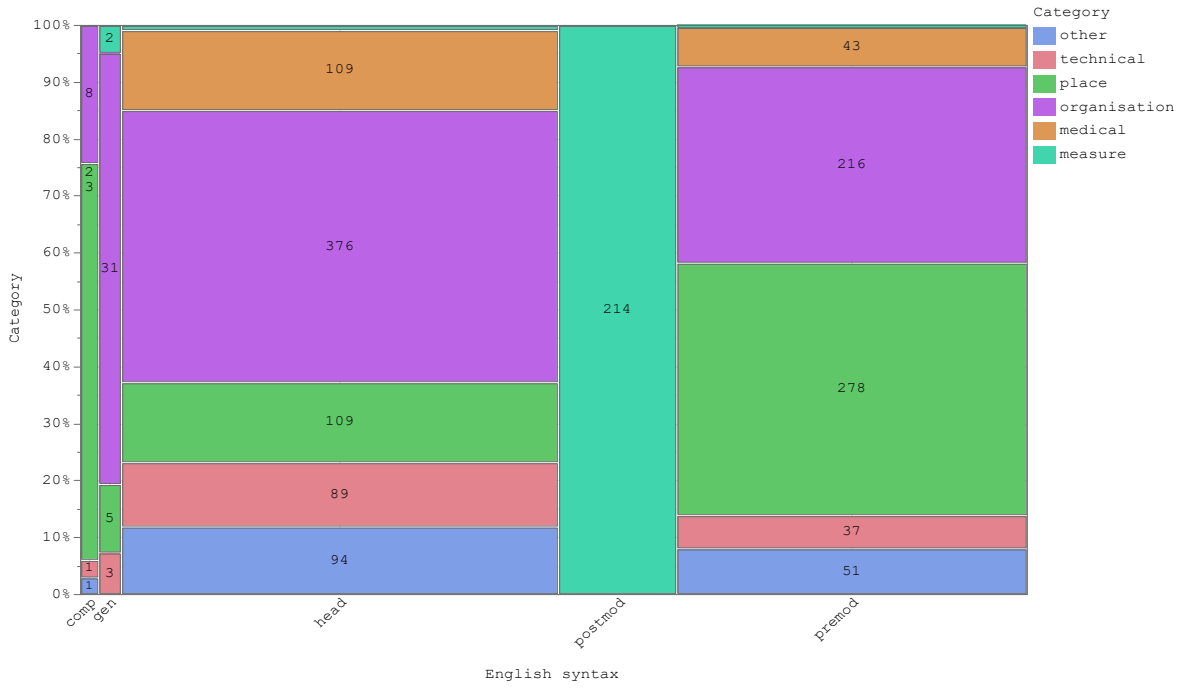

Figure 3. Syntactic functions and semantic categories of acronyms in English originals in LEGS.

Organisations are more strongly associated with heads (e.g., *When IBM introduced…*) than the place category, which in turn is more strongly associated with the premodifying function. However, the predominance of organisations among heads is much stronger than the predominance of place names among premodifiers. The differences between heads and premodifiers are partly explained by the highly frequent *US* and *UK*, which are typically used as premodifiers (e.g., *U.S. billionaires*), and partly by organisations also being rather frequent as premodifiers (e.g., *FBI agents*; *the former ICTY prosecutor*). From the frequent use of acronyms as premodifiers, it is evident that English writers readily exploit the syntactic

[7] As noted by one reviewer, both the category premodifier and genitive are in pre-head position, but due to their different forms and functions we keep them separated.

flexibility of the condensed acronyms (Fleischer, 1997: 189). Finally, as indicated above, the category of postmodifiers only comprises acronyms of measurement (e.g., *in the 50s CE*).

So far, the results have focused mainly on the LEGS source texts. In the following, the German and Swedish structural correspondences will be correlated with the originals. The findings shed light both on the translation process and language-specific tendencies.

4.3 German and Swedish correspondence types

The most notable finding is that about two-thirds of the English acronyms are kept in the German and Swedish translations.[8] The remaining third contains correspondences that lack an acronym altogether, instead being replaced by a spelt-out version or semantic equivalents, as will be described below.

In the German and Swedish translations, we identified nine different correspondence types. Most of these involve retaining an acronym in some form, while others rephrase the acronym in some way. First of all, (9) below exemplifies the use of acronyms as noun-phrase heads, a syntactic function that is quite frequent in translations (as also in the source language). Example (10) illustrates acronyms occurring as parts of German and Swedish hyphenated compounds (cf. Ström Herold and Levin, 2019). Similarly, Izwaini (2005: 85–86) proposes that the complex nature of English noun phrases with premodifying acronyms lead to them often being directly translated into Swedish (e.g., *OLE DB consumer* > *OLE DB-konsument*). Other categories are less frequent, such as (11) which illustrates the rare usage of acronyms in the genitive in translations. Target-language postmodifiers, given in (12), are also rare and only used to render English postmodifiers. A small number of acronyms are borrowed as premodifiers as parts of names as in (13).

Head

(9) a. According to the *FBI* (EN original)
b. Laut *FBI* (GE translation)
c. Enligt *FBI* (SW translation)

Compound

(10) a. the *fMRI* scanner (EN original)
b. einem *fMRT*-Gerät (GE translation)
c. en *fMRI*-skanner (SW translation)

Genitive

(11) a. he *NKVD's* interrogation system (EN original)
b. das Verhörsystem *des* [gen.] *NKWD* (GE translation)
c. *NKVD:s* [gen.] förhörsväsen (SW translation)

Postmodifier

(12) a. about 2500 *BCE* (EN original)
b. Omkring 2500 *f.Kr.* (SW translation)

Premodifier

(13) a. the battleship *HMS* Royal Oak (EN original)
b. das Schlachtschiff „*HMS* Royal Oak" (GE translation)

Apart from these five types that occur in both originals and translation, we identified four additional correspondence types that are exclusive to the translations: 1) semantic equivalents,

[8] Of the 1,699 English instances, 1,127 (66%) are rendered as acronyms in German and 1,147 (68%) in Swedish.

2) spell-outs, 3) prepositional phrases, and 4) omissions. The instances classified as semantic equivalents involve cases where translators have used conventionalised German and Swedish equivalents which are not acronyms, a strategy also noted by Ingo (2007: 121). This is exemplified in (14) by the English *NCOs* (short for *non-commissioned officers*) and its established Swedish non-acronym correspondent *underofficerare*. Spell-out refers to cases where the translations use the full underlying form of the acronym. This is illustrated in (15) where the German correspondence *Bruttosozialprodukt*[9] is the equivalent of the English acronym. The key difference between semantic equivalent and spell-out is that spell-outs consist of the full form of an acronym, while semantic equivalents are generalised, typically more culture-independent, term correspondents not related to the constituent parts of an acronym.

Semantic equivalent

(14) a. Recruits were constantly insulted and beaten by their *NCOs* (EN original)
b. *Underofficerarna* förolämpade och misshandlade ständigt rekryterna (SW transl.)
"under-officers"

Spell-out

(15) a. Nauru's entire *GDP* (EN original)
b. das *Bruttosozialprodukt* Naurus (GE translation)
"Nauru's Gross Domestic Product"

The two remaining translation correspondence types not attested in the source texts are paraphrases with prepositional phrases and omissions. A translation into a postmodifying prepositional phrase is given in (16). In omissions, as in (17), all information regarding the acronym is lost in the translation.

Prepositional phrase (PP)

(16) a. under strict *IAEA* supervision (EN original)
b. under strikt övervakning av *IAEA* (SW translation)
"supervision by IAEA"

Omission

(17) a. Similar shell middens exist all over the world from the *UK* to Australia, […]. (EN original)
b. Ähnliche Schalenhaufen gibt es überall auf der Welt Ø, […]. (GE translation)
"all over the world Ø"

The correspondence types show both differences and similarities in their distributions across the German and Swedish target texts. As illustrated in the radar plot in Figure 4 below, the main difference relates to compounds and to a lesser extent noun-phrase heads, semantic equivalents and spell-outs.[10]

[9] According to duden.de there is a German acronym, *BSP*, for this compound noun, but searches in the the DWDS corpus (dwds.de) indicate that it is not in regular use.

[10] Given the shared inventory of available structures in both target languages, we treat the adopted translation correspondence types as a classification problem and use the Kappa coefficient to assess symmetry; 0 indicates complete lack of agreement and 1 indicates complete agreement. The overall Kappa coefficient for agreement across the whole table shows moderate symmetry (K=0.48, se=0.013). Calculating Kappa for each target-language structure, we get the order from highest to lowest as postmodifier (K=0.74, se=0.05), head (K=0.40, se=0.02), compound (K=0.320, se=0.106), genitive (K=0.27, se=0.013), and premodifier (K=0.21, se=0.022).

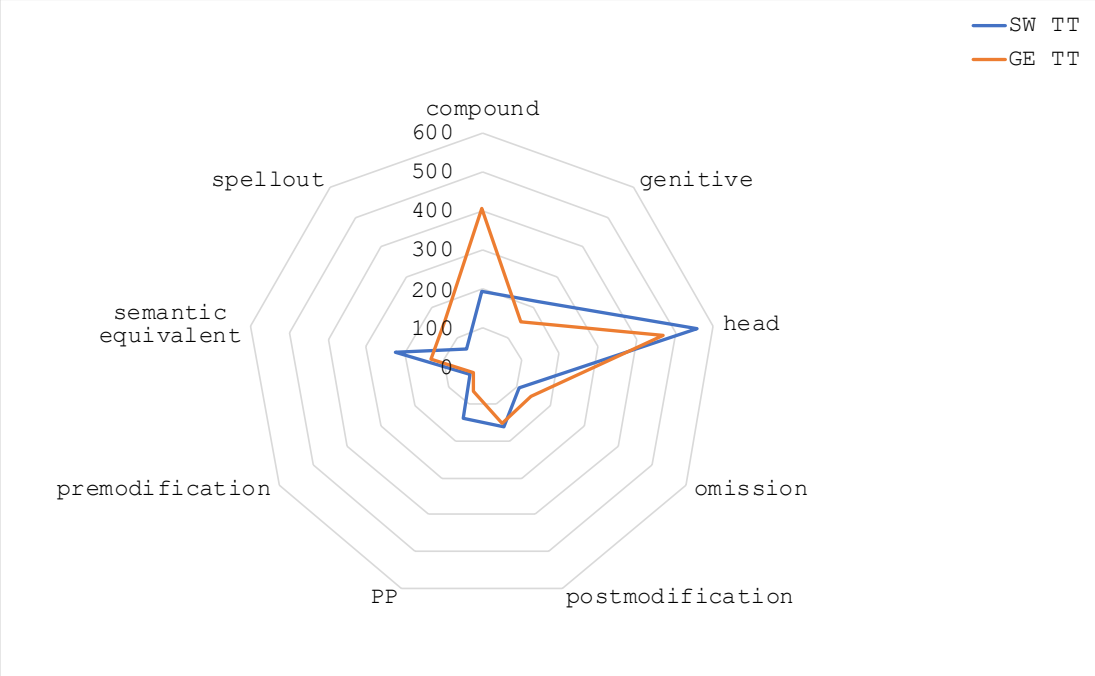

Figure 4. Distributions of correspondence types in German and Swedish translations in LEGS.

The stronger German preference for compounds was also found in Carlsson's investigation (2004: 75, 138) of German and Swedish newspaper language, and also Ström Herold and Levin's (2019) study on English proper noun premodifiers in German and Swedish translation. In contrast to the German compound affinity, Swedish more heavily relies on noun-phrase heads and semantic equivalents. In search for explanations for the target-language differences seen above, we divided all instances according to the semantic categories presented in Table 1 (measure, medical, organisation, place, technical and other) and the translation correspondences. The results are presented in Figure 5.

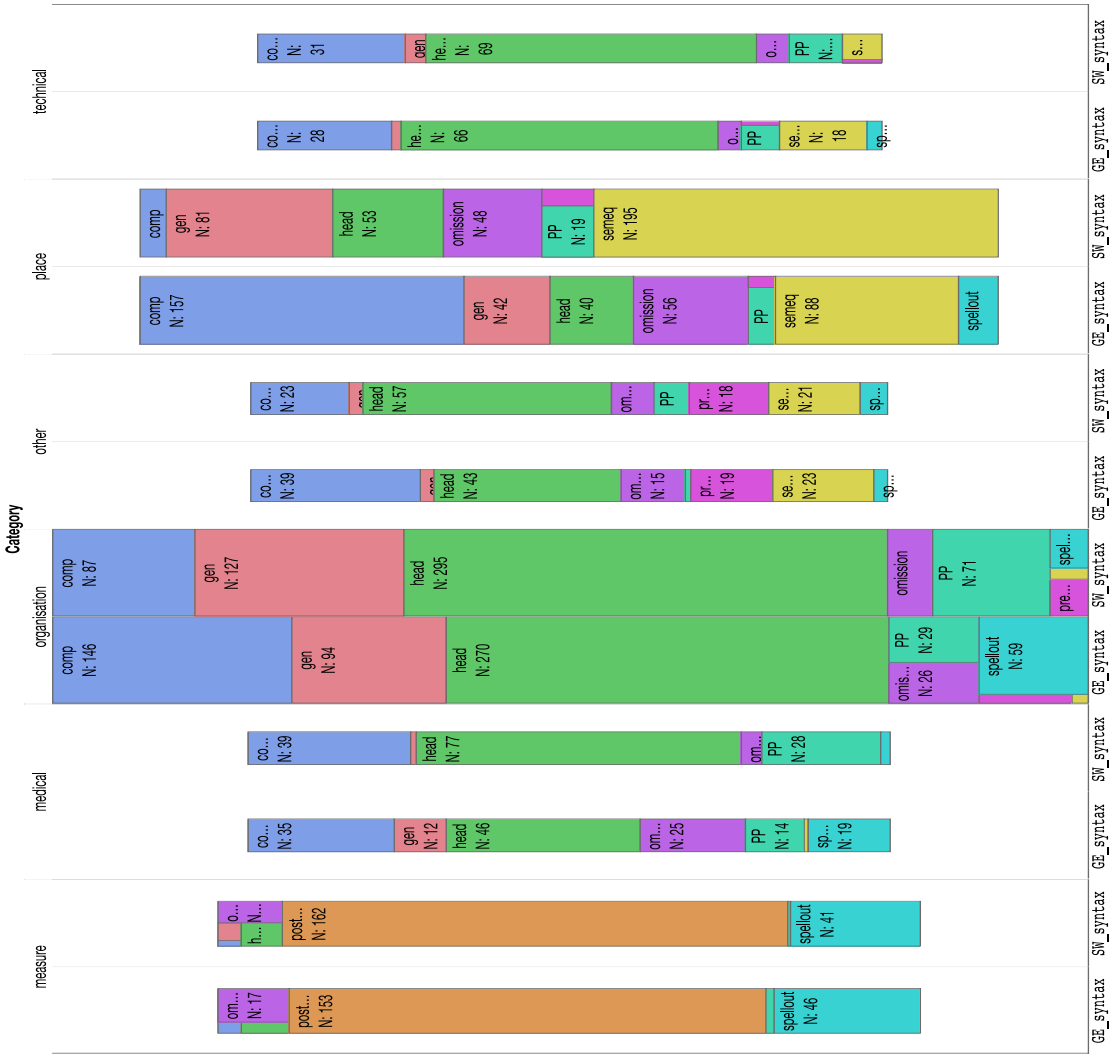

Figure 5. Distributions of correspondence types in German and Swedish translations by semantic category in LEGS.

The figure shows that the main differences between German and Swedish relate to organisations and places. German compounds are particularly frequent with acronyms referring to places and organisations, one strong factor being the frequent compounds with *US* (e.g., *der US-Comedian*; *US-Politiker*). Organisation name compounds also occur in German (e.g., *NATO bombing* > *NATO-Bombardement* (GE); but cf. *Natos bombkampanj* (SW) 'NATO's bombing campaign'), but to a lesser extent. The Swedish predilection for semantic equivalents is partly the reverse of the German *US*- compounds, as many of these involve the adjective *amerikansk* for English *US* (e.g., *den amerikanska komikern* 'the American comedian'), the translation option suggested by Magnusson (1987: 91). The slightly greater preference for spell-outs in German translations may be a reflection of a general tendency in our material for German translators to use more explicit correspondences than Swedish translators. This was exemplified above in (1) where the English acronym *RAF* was spelt out by the German translator while the Swedish translator opted for the acronym only. In other cases, the German translations contain translated spell-outs while Swedish retains the English acronym, as in the medical example *the PFC* > *der präfrontale Kortex* (GE); *PFC* (SW).

Thus far the focus has been structural preferences in originals and target texts. The findings regarding correspondence types, in particular spell-outs, have also touched upon the degree of explicitness in translation. This theme will be explored further in section 4.4.

4.4 Acronyms and explicitation

As discussed above, acronyms may be highly culture-specific (Ingo, 2007: 123), and consequently readers of translations cannot always be expected to be familiar with them. In such cases, translators have a range of options at their disposal, many of which are more explicit than the original expressions. Section 4.4.1 discusses how and to what extent acronyms are introduced and explained in translations, and 4.4.2 focuses on language choice in these explicitations.

4.4.1 *Introducing and explaining acronyms*

To facilitate comprehension, translators may opt to insert explanations with different degrees of explicitness (see, e.g., Blum-Kulka, 2004 [1986]). In (18), the German translator adds a contextual clue, the hypernymic descriptor *Studierfähigkeitstests*, putting the acronym *SAT*[11] in brackets. This is an efficient and unobtrusive way for a translator to enhance readability.

(18) a. […] their children's *SAT* verbal and quantitative scores, […]. (EN original)
b. […] die Punktwerte ihrer Kinder im verbalen und mathematischen Teil *des Studierfähigkeitstests (SAT)*. (GE translation)
"study-aptitude-test.GEN"

In other cases, a contextual clue is already given in the original which is then transferred to the translation. This is seen in (19), where *U.S.* gives rise to *amerikanischen* in the German translation.

(19) a. In 2007, *the three major U.S. networks – CBS, NBC,* and *ABC* – ran 147 stories on climate change. (EN original)
b. 2007 brachten *die drei großen amerikanischen Fernsehgesellschaften – CBS, NBC* und *ABC* – 147 Beiträge über den Klimawandel. (GE translation)
"the three big American TV-companies"

Although the cultural distances between the Anglophone world and Germany and Sweden may be surmised to be relatively small, the LEGS data reveal significant differences in explanation practices in German and Swedish translations. In general, German translators explain acronyms more often than Swedish ones and they do so predominantly in German, while, in comparison, Swedish translators use more English in their explanations. These tendencies are exemplified in (20):

(20) a. But another aspect […] has been […] surrendered to *the United States National Security Agency (NSA)* […]. (EN original)
b. Darüber hinaus wurde […] ein weiterer Aspekt […] an *die US-amerikanische Nationale Sicherheitsagentur (NSA)* abgetreten, […]. (GE translation)
"the American national security-agency (NSA)"
c. Men ännu en aspekt […] har […] överlämnats till *USA:s National Security Agency (NSA)* […]. (SW translation)

If we consider instances where there is no explanation provided in the English original, such as a descriptor introducing the acronym as in (19), we find 209 added explanations in the German translations as opposed to only 95 in the Swedish. This difference is highly

[11] Acronym for *Scholastic Aptitude Test*.

significant.[12] The larger proportion of explicitation (Blum-Kulka, 2004 [1986]) in German translation is due to German readers being less likely to be familiar with the English language and Anglophone culture than Swedish readers are.[13] The overall inclination for German translators to avoid English more than Swedish ones might also be related to the differences in status of the languages. The status of German is higher than Swedish, as reflected in more texts being translated from the former language (cf. UNESCO's *Index Translationum*), and thus German translators seem to "dare" to introduce more changes in translations than Swedish ones do (Levin and Ström Herold, this volume).

The following examples illustrate the strategy of adding target-language explanations, sometimes in both translations and sometimes in only one. The target-language explanation can be a more or less direct translation of the original English full form, as in (21) where the English acronym *RSPB* (for *The Royal Society for the Protection of Birds*) is explained using the respective target languages, or a more descriptive paraphrase, as in the added German apposition in (22b). In (22c), the Swedish translator transfers the source-text acronym with no additional explanation.

(21) a. […] one which had been developed by the *RSPB* for monitoring birds' nests. (EN original)
b. […] ein von der *RSPB (Königliche Gesellschaft für Vogelschutz)* entwickeltes System zur Beobachtung von Vogelnestern. (GE translation)
"royal society for bird-protection"
c. […] ett som hade utvecklats av *RSPB (Kungliga fågelskyddssällskapet)* för att övervaka fågelbon. (SW translation)
"royal bird-protection-society"

(22) a. In 1990, the *NSPCC* estimated there were 7,000 known images of child pornography in circulation. (EN original)
b. 1990 schätzte die *NSPCC, ein britischer Kinderschutzverein*, die Zahl der in Umlauf befindlichen Fotos mit Kinderpornografie auf 7.000. (GE translation)
"a British child-protection-agency"
c. År 1990 uppskattade *NSPCC* att det fanns 7000 kända barnpornografiska bilder i omlopp. (SW translation)

The correlations between the semantic categories of the acronyms and the likelihood of translators furnishing them with explanations in the target texts are given in Table 2.

Table 2. German and Swedish explanation likelihood by semantic category (* denotes a statistically significant difference between German and Swedish TTs for that semantic category).

	German explanation				**Swedish explanation**			
	no		**yes**		**no**		**yes**	
Semantic category	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
Measure	80	180	20	45	80.89	182	19.11	43
Medical*	69.74	106	30.26	46	89.47	136	10.53	16
Organisation*	78.13	493	21.87	138	85.58	540	14.42	91
Other	66.44	97	33.56	49	72.6	106	27.4	40
Place*	93.73	389	6.27	26	100	415	0	0
Technical	90	117	10	13	90.77	118	9.23	12

[12] χ^2=46.4, df=1, p=***

[13] See, e.g., the *First European Survey on Language Competences: Final Report* (2012) where Swedish pupils' English skills were the highest in all the countries surveyed.

German and Swedish translations are quite similar when it comes to explaining measure acronyms, technical acronyms and acronyms of the class 'other', but there is a significant preference for explanations in German translations with medical acronyms, organisation acronyms and place acronyms.[14] These trends will be discussed and exemplified in the next section.

4.4.2 *Language choice in explicitations*

Based on our data, we further classified the explanations into four different subtypes (apart from no explanation) based on the language(s) the explanation is written in: i) English, ii) target language, iii) target language with a contextual clue, and, finally, iv) mixed languages, meaning that both English and the target language are used in the explanation part. These different explanation strategies will be discussed in more detail below, but first a quantitative overview in Table 3:

Table 3. Language choice in explanations by semantic category (* denotes a statistically significant difference between German and Swedish TTs for that semantic category).[15]

	Semantic category					
German TT	**measure**	**medical***	**organisation***	**other**	**place***	**technical**
English	0	2	49	2	0	1
mixed languages	1	1	4	9	0	2
no explanation	180	106	493	97	389	117
target language	44	27	73	29	20	4
target language + contextual cue	0	16	12	9	6	6
Swedish TT						
English	1	4	43	7	0	1
mixed languages	0	1	8	5	0	1
no explanation	182	136	540	106	415	118
target language	41	8	28	19	0	4
target language + contextual cue	1	3	12	9	0	6

Looking at the different ways of explaining the acronyms, we see that the strategies are largely similar in German and Swedish, with the use of English explanations and mixed languages being substantially the same. In both the German and Swedish translations explanations in English are predominantly used for organisation acronyms. Notably, German translations contain nearly three times more target-language explanations of organisation acronyms than Swedish translations. Looking closer, however, it becomes apparent that this observation is

[14] The independence of the choice of explication type was tested for each semantic category using Pearson's chi-squared test and the effect size using phi; in the present study we consider each instance of translation as an independent occurrence. The significance levels were: measure (χ^2=0.01, df=1, p=ns), medical (χ^2=14.9, df=1, p=***, phi=0.23), organisation (χ^2=11.29, df=1, p=***, phi=0.09), other (χ^2=1.03, df=1, p=ns), place (Fisher's exact p=***, phi=0.17), and technical (χ^2=0, df=1, p=ns).

[15] The independence of the choice of explication type was tested for each semantic category using Pearson's chi-squared test except for measure and place, for which Fisher's exact test was used due to cell counts of zero; the effect size is expressed as Cramér's V. The significance levels were: measure (Fisher's p=ns), medical (χ^2=23.6 df=4, p=***, V=0.27), organisation (χ^2=23.9, df=4, p=***, V=0.13), other (χ^2=6.4, df=4, p=ns), place (Fisher's exact p=***, V=0.17), and technical (χ^2=0.3, df=4, p=ns).

somewhat misleading, because 22 out of the 73 occurrences are translations of the same acronym, *BEF* (for *British Expeditionary Force*), exemplified in (23).

(23) a. […] he wasted no time in turning his attention back to the war and the advance of the *BEF* into Belgium. (EN original)
b. […] wandte er sich, ohne Zeit zu verlieren, wieder dem Krieg und dem Vormarsch *des Britischen Expeditionskorps* nach Belgien zu. (GE translation)
"the British expeditionary-corps"

Similarly, 13 out of 26 occurrences of the medical acronym *EF* (*Executive Function*) are spelt-out in German:

(24) a. Children need *EF* to resist temptations beyond marshmallows […]. (EN original)
b. Kinder benötigen *die Exekutiven Funktionen*, um auch anderen Versuchungen als Marshmallows zu widerstehen […]. (GE translation)
"the executive functions"

These cases often involve examples where the English original includes a spell-out, i.e. a full form of the acronym which is directly transferred into both translations without further explanation:

(25) a. *The Internet Watch Foundation (IWF)* is a UK-based organization […]. (EN original)
b. *Die Internet Watch Foundation (IWF)* ist eine Organisation mit Sitz in Großbritannien […]. (GE translation)
c. *Internet Watch Foundation (IWF)* är en organisation med bas i Storbritannien […]. (SW translation)

However, we also find cases where the translator adds a spelt-out English version of the acronym not present in the original. Many of these cases are culture-specific, as in the following example where the addition clarifies the meaning of the letters. It should be noted that the strategy presupposes some knowledge of English from the Swedish readers.

(26) a. […] supported by a wide range of religious groups but opposed by the *ACLU*. (EN original)
b. […] som stöddes av ett brett spektrum av religiösa grupper men motarbetades av *American Civil Liberties Union (ACLU)*. (SW translation)

Mixed-language explanations are much rarer than English explanations in both the German and Swedish translations, the German in (3b) above being one of the exceptions. Another highly explicit way of rendering the acronym is given in (27) below, where the Swedish translation stacks three different versions of the organisation name: in Swedish, spelt out in English and as an English acronym.

(27) a. Meanwhile, *the British Expeditionary Force (BEF)* was preparing its departure for France […]. (EN original)
b. Under tiden förberedde sig *brittiska expeditionsstyrkan, British Expeditionary Force (BEF)* […]. (SW translation)
"the British expeditionary-force"

This overly explicit and rather cumbersome translation is likely the result of two conflicting objectives: the translator's loyalty towards the source text and a wish to bring the source text closer to the new target-text readers. In this particular case, the acronym does not recur again in the Swedish translation and, thus, could be deemed to be superfluous, making it a candidate for omission.

As mentioned in connection with (19), target-language clues may have a correspondence in the English original, but they may also be added to the target text. The latter alternative is

seen in the German version in (28b) where the hypernym *Programmiersprache* has been added, while the Swedish translator adheres more closely to the English source text.

(28) a. He did a great version of *BASIC* […]. (EN original)
b. Er erstellte eine großartige Version der *Programmiersprache BASIC* […]. (GE translation)
"the programming language BASIC"
c. Han skrev en jättebra version av *BASIC* […]. (SW translation)

Finally, we will consider those exceptional cases where a translator reduces the degree of explicitness. Some of these depend on the source text being more explicit than may be deemed strictly necessary. One example is seen in (29), where the English original for the second time after several pages re-introduces the German acronym *OKH*, which stands for *Oberkommando des Heeres* ('the army high command'). The German translator here only retains the acronym while omitting the descriptive paraphrase. The fact that the acronym was spelt out previously – in both original and translation – and the fact that the acronym is likely to be more recognisable to the German target audience make the use of the bare acronym a feasible choice for both languages here.

(29) a. *The army high command, the OKH,* was instructed […]. (EN original)
b. *Das OKH* erhielt Weisung, […]. (GE translation)

However, the main observations in this section still hold true: German translators add more explanations than Swedish ones do, and they do so predominantly in their first language.

4.4.3 *Acronym frequency and explanations*

As discussed at the beginning of section 4, acronyms vary widely when it comes to how frequent they are in a language, and how generic or specialised they are in meaning. Intuitively, we would expect the less common and more specialised acronyms to require explicitation more than the common and generic ones.

To examine the relationship between an acronym's real-world frequency and the translators' strategy in our data, we obtained the frequencies of the acronyms from Google Books (UK) following the procedure introduced in section 3. Figure 6 shows the frequencies of the acronyms on a log10 scale.

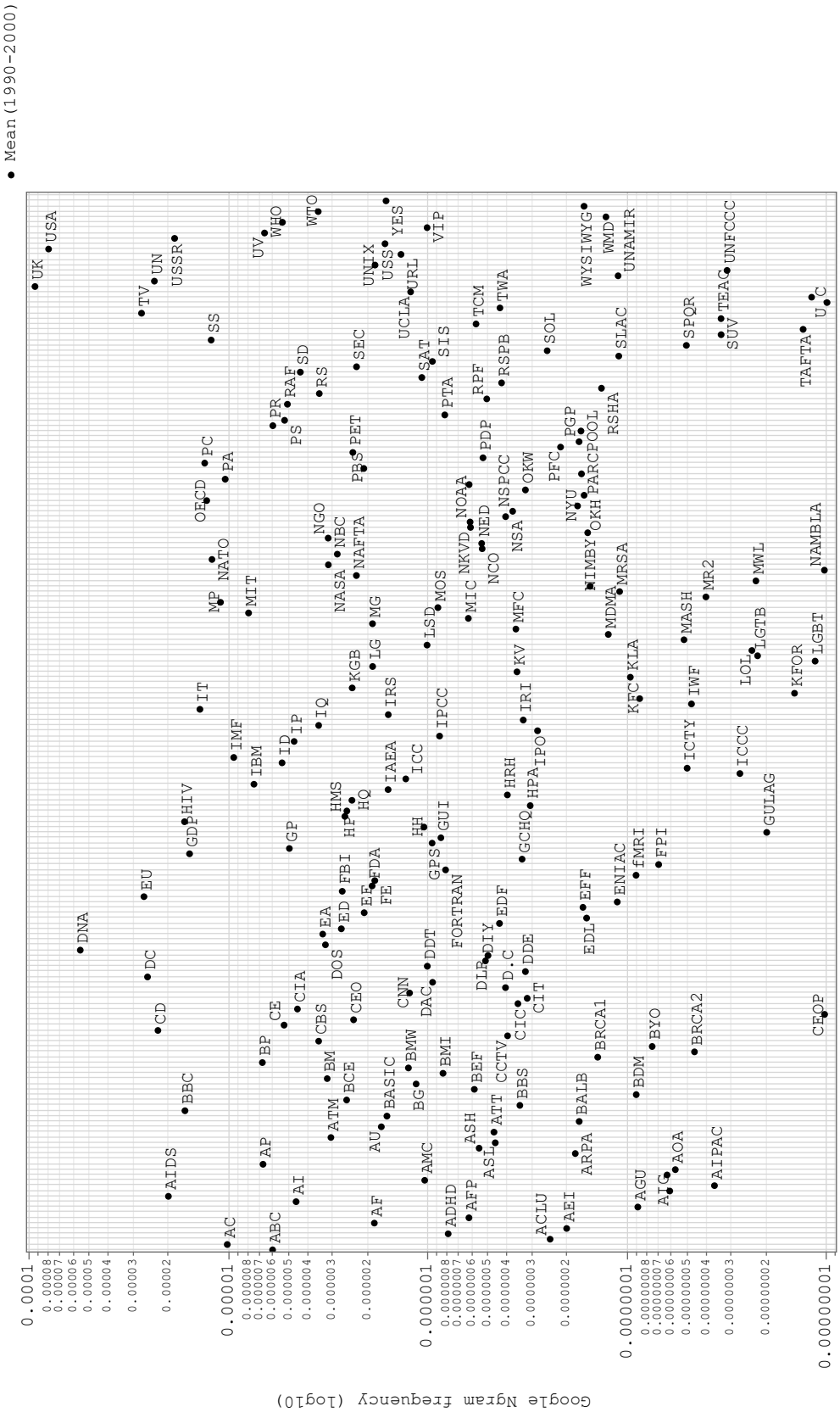

Figure 6. Frequencies in Google Books (British, 1990–2000) of the acronyms in LEGS.

The natural steps of the log10 scale can be used as a heuristic method for categorising the acronyms into three frequency bands. At the top in the first band, with frequencies ranging

from 0.0001 to 0.001, we find very common acronyms such as *DNA*, *BBC*, *NATO* and *OECD*. In the second band (0.00001 to 0.00001), we see *GP*, *IMF*, *NASA* and *WTO*, still acronyms that most mature competent readers would recognise. In the third band (0.0000001 to 0.000001), we find most of the medical and technical acronyms, which readers are increasingly unlikely to know unless they are previously familiar with the specific field. The three frequency bands were turned into a categorical variable with the levels COMMON, MODERATE and RARE.

Figure 7 shows that the proportion of instances where the translators explain the acronyms agrees with the hypothesis that rare items are more likely to be explained. The values show the number of acronyms that were explained and not explained; in some cases, the same acronym was explained several times, but multiple instances are conflated here simply as 'explanation'. The differences between the German and Swedish translators are not statistically significant in any of the bands.[16]

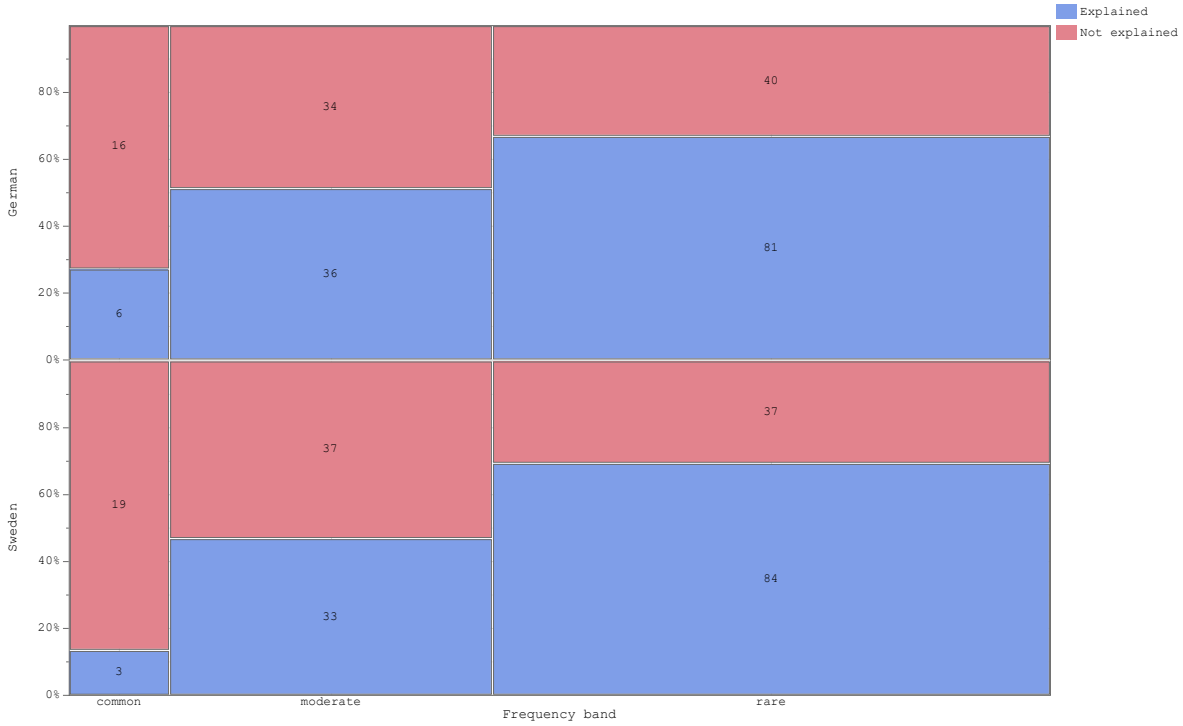

Figure 7. German and Swedish translators' choice to provide an explanation in relation to the frequency of the acronym in contemporary written English texts.

When we turn to the breakdown of explanation types based on the frequency bands of the acronym, we find a partly different picture, as seen in Figure 8:

[16] The independence of the choice between explanation and no explanation was tested using Pearson's chi-squared test. The results show no statistically significant differences between German and Swedish translations: common acronyms χ^2=0.56, df=1, p=ns; moderate acronyms χ^2=.011, df=1, p=ns; rare acronyms χ^2=3.51, df=1, p=ns.

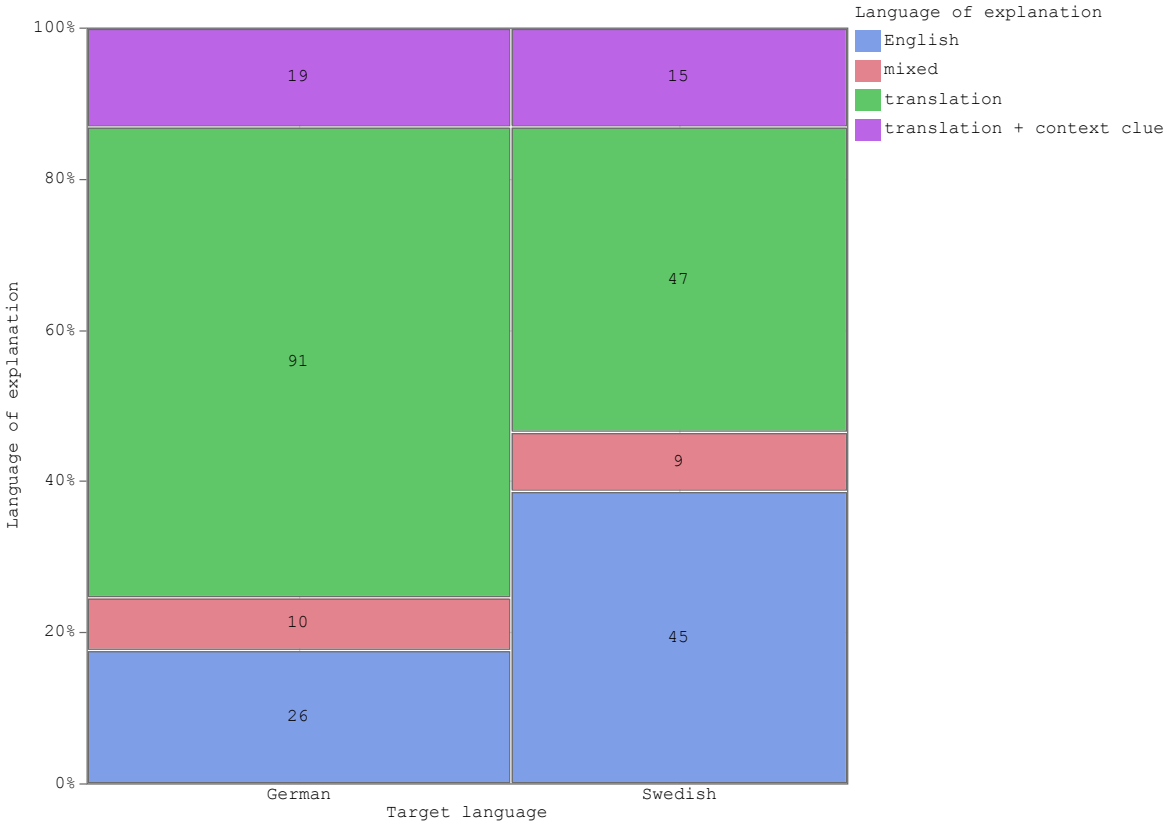

Figure 8. Language used in explanations of RARE acronyms.

German and Swedish translators use the different explanation types equally when it comes to COMMON and MODERATELY rare acronyms.[17] However, when it comes to RARE acronyms, Swedish translators show equal preference for translated explanations and for using English, while German translators show a clear preference for translations.[18]

This section has shown different techniques for introducing and explaining acronyms in texts. The original may already contain contextual clues, such as a hypernym introducing the acronym, which the translators can make use of, meaning that no additional explanations are necessary. In other cases a hypernym may be added by translators, combined with a direct transfer of the acronym, in order to facilitate comprehension. This section has also shown that the frequency of the English acronyms in general has some effect on explanation practices in translations.

5. Conclusions

The present study investigated how German and Swedish translators of English non-fiction texts approach acronyms. The primary data consist of 1,699 acronyms retrieved from the trilingual translation corpus LEGS. The acronyms were categorised into five main semantic categories based on their domain of use. The relative sizes of these categories verified earlier findings by Leech *et al.* (2009).

The first research question concerned the syntactic functions that acronyms fulfil in the source and target texts. The three languages all belong to the Germanic family of languages, which means that similar syntactic functions are available in all three languages. A cross-

[17] Common acronyms χ^2=0.07, df=3, p=ns; moderate acronyms χ^2=2.52, df=3, p=ns.
[18] Rare acronyms χ^2=16.42, df=3, p=***, Cramér's V=0.25.

tabulation of semantic categories and syntactic functions in the source texts revealed substantial correlations, the most notable being organisation acronyms frequently being used as noun phrase heads while place acronyms more often are used as premodifiers.

Turning to the target texts, the same functions were observed in both translation subcorpora, allowing symmetrical comparison between the two. The data reveal that German and Swedish translators largely rely on the same set of correspondence types, but some language-specific differences were also observed. In particular, the German translators favour compounds more than the Swedish translators (cf. Ström Herold and Levin, 2019), while the Swedish translators are slightly more inclined to using noun phrase heads and prepositional phrases as correspondences. It would be fruitful to perform a follow-up study of German and Swedish source texts as this may reveal language-specific preferences for, e.g., acronyms in compounds. This may in turn explain some of the function-related differences observed here.

The second research question focused on explanation practices in the translations. The German translators include clarifying explanations somewhat more than their Swedish counterparts − 22% against 13% − with the breakdown being more or less similar across the semantic categories. We also analysed the language choices of these explicitation strategies, observing that the use of the target language was the primary preference in both languages. A notable finding is nevertheless that the preference for using the target language in German is even stronger than in Swedish, which instead incorporates more English material.

Finally, the third research question addressed the extent to which the frequency of the acronyms in contemporary English might explain the need for explicitation. A comparison with frequencies in Google Books predictably showed that rare acronyms are explained more often than moderately common or common acronyms, with the German translators showing strong preference for translations while the Swedish translators also used explanations in English to a notable extent.

The overall findings of the study show that German and Swedish translators largely use similar strategies when translating acronyms. However, there were also some significant differences, which may at least in part be explained by how familiar German and Swedish readers are expected to be with English acronyms. It is also likely that the status differences between the languages play a role here (see UNESCO's *Index Translationum*). Regarding language choice, which was a prominent feature in this study of acronyms, a broader investigation on multi-lingual practices in texts would be a welcome contribution in the future. What kinds of foreign elements are included, adapted or translated in both originals and translations? Another avenue of acronym research could more strongly emphasise the contrastive aspect by comparing practices in originals to determine if there are differences in how languages introduce acronyms in texts, or if there are universal strategies.

References

Betancourt Ynfiesta, B., Treto Suárez, L. and Fernández Peraza, A.V. 2013. Translation of Acronyms and Initialisms in Medical Texts on Cardiology. *CorSalud* 5(1): 93–100.

Blum-Kulka, S. 2004 [1986]. Shifts of Cohesion and Coherence in Translation. In *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition*, J. House and S. Blum-Kulka (eds), 17–35. Tübingen: Narr.

Carlsson, M. 2004. *Deutsch und Schwedisch im Kontrast: Zur Distribution nominaler und verbaler Ausdrucksweise in Zeitungstexten*. PhD dissertation, Gothenburg University.

Ehrmann, M., Della Rocca, L., Steinberger, R. and Tannev, H. 2013. Acronym Recognition and Processing in 22 Languages. *Proceedings of recent advances in natural language processing*, 237–244.

First European Survey on Language Competences: Final Report. 2012. Accessed 23 September, 2019. Available at: https://crell.jrc.ec.europa.eu/?q=article/eslc-database [last accessed 12 November 2020].
Fleischer, W. 1997. *Phraseologie der deutschen Gegenwartssprache*. Tübingen: Niemeyer.
Fleischer, W. and Barz, I. 2012. *Wortbildung der deutschen Gegenwartssprache*. Berlin: De Gruyter.
Gale. 2007. *Acronyms, Initialisms, and Abbreviations Dictionary*. 38th ed.
Giegerich, H.J. 2004. Compound or Phrase? English Noun-plus-noun Constructions and the Stress Criterion. *English Language and Linguistics* 8(1), 1–24.
Index Translationum. UNESCO. N.d. https://en.wikipedia.org/wiki/Index_Translationum. [Last accessed 22 June 2021]
Ingo, R. 2007. *Konsten att översätta. Översättandets praktik och didaktik*. Lund: Studentlitteratur.
Izwaini, S. 2005. Corpus-based Study of IT Terms. *ESP Across Cultures* 2, 76–93.
Kobler-Trill, D. 1994. *Das Kurzwort im Deutschen. Eine Untersuchung zu Definition, Typologie und Entwicklung*. Tübingen: Max Niemeyer Verlag.
Leech, G., Hundt, M., Mair, C. and Smith, N. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge: Cambridge University Press.
Levin, M. and Ström Herold, J. This volume. On Brackets in Translation (or How to Elaborate in Brackets). *Bergen Language and Linguistics Studies* 11(1), 120–143.
Magnusson, G. 1987. *Från tyska till svenska. Översättningsproblem i sakprosa*. Malmö: Liber.
Mair, C. 2006. *Twentieth-century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.
Newmark, P. 1988. *A Textbook of Translation*. New York: Prentice Hall.
Nuopponen, A. and Pilke, N. 2016 [2010]. *Ordning och reda. Terminologilära i teori och praktik*. Lund: Studentlitteratur.
Nübling, D. and Duke, J. 2007. Kürze im Wortschatz skandinavischer Sprachen. Kurzwörter im Schwedischen, Dänischen, Norwegischen und Isländischen. In *Sprachliche Kürze. Konzeptuelle, strukturelle, und pragmatische Aspekte*, J. A. Bär, T. Roelcke and A. Steinhauer (eds), 227–263. Berlin/New York: de Gruyter.
Pechenick, E.A., Danforth, C.M. and Dodds, P.S. 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS ONE 10*(10).
Sigurd, B. 1979. Förkortningarna och det moderna samhället. *Språkvård. Tidskrift utgiven av Svenska språknämnden* 2, 3–8.
Ström Herold, J. and Levin, M. 2018. English Supplementive *ing*-clauses and their German and Swedish Correspondences. *Corpora et Comparatio Linguarum: Textual and Contextual Perspectives*, S.O. Ebeling and H. Hasselgård (eds) *Bergen Language and Linguistics Studies* 9(1): 115–138.
Ström Herold, J. and Levin, M. 2019. *The Obama Presidency*, *the Macintosh Keyboard* and *the Norway Fiasco*: English Proper Noun Modifiers in German and Swedish Contrast. *English Language and Linguistics* 23(4): 827–854
Xu, H., Stetson, P.D. and Friedman, C. 2007. A Study of Abbreviations in Clinical Notes. In *AMIA annual symposium proceedings* (American Medical Informatics Association), 821–825.

Authors' addresses

Jenny Ström Herold / Magnus Levin / Jukka Tyrkkö
Linnaeus University
Department of Languages
SE-351 95 Växjö
Sweden
jenny.strom.herold@lnu.se / magnus.levin@lnu.se / jukka.tyrkko@lnu.se

Eyes and speech in English, Finnish and Czech children's literature[1]

> The fires of fury and hatred were smouldering in her small black eyes.
> "Matilda!" she barked.
> "Stand up!"
>
> [Roald Dahl, *Matilda*, 1988]

Anna Čermáková, Markéta Malá

Charles University, Prague (Czech Republic)

This study explores cross-linguistically, in English, Czech and Finnish, eye-behaviour that occurs in children's fiction in the vicinity of character speech. We explore how authentic eye behaviour, as an important part of non-verbal communication, is rendered in fictional worlds. While there are more similarities than differences across the languages in the characteristics and narrative functions of fictional eye-behaviour, the linguistic encoding differs substantially due to typological differences between the languages. The same semantic roles are often expressed by divergent syntactic means. The divergence is reflected primarily in the relative weight of different word-order principles, the different means of indicating simultaneity, as well as the role of inflection in Finnish and Czech.

Keywords: fictional speech, eye-behaviour, gaze, children's literature, language typology, Czech/English/Finnish

1. Introduction

The description of the scene in the quote above from Roald Dahl's iconic text, where the smouldering fires of fury and hatred in Miss Trunchbull's small black eyes as she barks at Matilda to stand up allows us to fully immerse in the moment of confrontation between Miss Trunchbull and Matilda. The tense atmosphere is created with the use of a very short emphatic direct speech graphically emphasised by the use of exclamation marks. The direct speech is introduced by a body language description with a focus on eye-behaviour, then it is interrupted by an expressive reporting clause (*she barked*), which suggests that the speech is loud and

[1] This research was supported by the European Regional Development Fund project 'Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World' (reg. no.: CZ.02.1.01/0.0/0.0/16_019/0000734), the programme 'Progres Q08 *Czech National Corpus*, 'Progres Q10 Language in the shiftings of time, space, and culture', and 'Progres Q17 The Teachers Preparation and Profession in the Context of Science and Research' implemented at Charles University.

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.v11i1.3444.

aggressive. As readers, we can deduce a great deal of information about the event itself and the character of Miss Trunchbull.

This study is focused on fictional body language, and more specifically eye-behaviour that occurs in connection with fictional speech as illustrated in the above extract. Descriptions of body language in fiction are not only an important part of characterisation but they also refer to the physical body of fictional characters and describe how fictional people relate to each other. In fictional texts, body language descriptions play a particularly important role in connection with speech because they contribute to the effect of its authenticity. While the authenticity of fictional speech has been studied extensively (see, for example, Mahlberg *et al*., 2019), we will, drawing on Argyle's (2010) framework of bodily communication, examine how the fictional eye-behaviour resembles the authentic and what linguistic means typologically different languages use to encode the eye-behaviour in fiction.

The study is a contrastive one. We will examine fictional eye-behaviour across three languages: English, Czech and Finnish. While we can hypothesise there will not be substantial cultural differences in the types of eye-behaviour described because all three languages belong to low-contact cultures (Argyle, 2010), we can expect substantial linguistic differences due to the different language typologies represented. English is a predominantly analytic Germanic language with fixed word order, Czech is a West Slavic inflectional language with free word order and Finnish is an agglutinative Finno-Ugric language.

The study relies on data from comparable corpora of non-translated children's fiction. Children's literature is a specific text-type in several respects. Its intended readers are only gradually developing their reading and cognitive skills, and also real-life knowledge – Nikolajeva (2014) refers to them as 'novice readers'. It is therefore expected that the linguistic make-up of these texts will reflect the readership. The language of children's literature has received surprisingly little attention (see, for example, Stephens, 2004; Wild *et al*., 2013) and even less so cross-linguistically (but see, for example, Čermáková and Chlumská, 2017).[2] One of the features that has been expected and observed is a greater degree of explicitness than in texts written for adult readers (Šebestová and Malá, 2019). While fictional body language is difficult to describe systematically because of its variety of forms (see Mahlberg *et al*., 2020; Čermáková and Mahlberg, forthcoming), we assume that children's fiction includes eye-behaviour descriptions that are accessible to the 'novice reader' and will thus constitute a suitable data-source for mapping this phenomenon cross-linguistically.

'Eyes' are one of the most frequently mentioned body parts in fictional texts and looking, or 'gaze', "is of central importance in social behaviour" (Argyle, 2010: 153). The eyes are also one of the most important channels for the expression of emotions (ibid.: 5). Understanding norms of social behaviour and encoding of emotions are of crucial importance for a child's development. Children's fiction is one source that encodes these norms. The meaning making process, though it may be empirically "problematic to access and judge readers' cognitive and emotional engagement with texts" (Nikolajeva, 2014: 2), is based on the interaction between the information in the text and the reader's real-life knowledge. However, so far, less attention has been paid to "the profound difference between young and adult readers" (ibid.: 10). Nikolajeva explicitly inquires what happens when the "readers' capacity to engage with texts is absent or underdeveloped" and she asks "how texts may deliberately compensate for this obstacle" (ibid.), that is, what the meaning-making process of children is like and in what ways, if at all, children's texts support it.

In view of the above, we aim to answer the following research questions:

[2] Translation of children's literature has received comparatively more attention, see, for example, Alvstad (2010), Lathey (2011), Čermáková and Mahlberg (2018).

1. What are the lexico-grammatical similarities and differences in fictional eye-behaviour descriptions in the typologically different languages: English, Czech and Finnish?

2. What are the characteristics and narrative functions of fictional eye-behaviour descriptions across the three languages?

In Section 2, we describe the theoretical background of this study and in Section 3, the data and methodology used. Section 4 aims to answer our first research question and looks at lexico-grammatical similarities and differences across the three languages. Section 5 aims to answer our second research question and looks at the characteristics and narrative functions of fictional eye-behaviour. Section 6 offers conclusions and suggestions for further study.

2. Body language and speech

The centrality of character in fiction has been recognised for some time now; Stockwell and Mahlberg (2015: 130) suggest that "the relationship that readers develop with fictional characters is a main motivating factor in reading literature at all". We can assume that communication between characters is key for meaning making. Communication between real people is also something that children are exposed to daily and are learning to make sense of. The "meaningfulness" of fictional communication depends on "representing the kind of language which a reader can recognise, by observation, as being characteristic of a particular situation" (Leech and Short, 2007: 129).

The representation of characters' speech as part of characterisation has received considerable attention (Leech and Short, 2007; Semino and Short, 2004). One of the concerns has been the credibility and authenticity of its representation (McIntyre, 2016). While the overlap between fictional speech and "real" speech still lacks a large-scale systematic analysis (but see Mahlberg *et al*., 2019), Page (1988: 7–10) points to inherent characteristics of spoken language, such as pauses, repetitions, grammatical inconsistencies, its dependence on the shared context and the "phonological component" that make it difficult to adequately and meaningfully re-create in fictional writing. Some of these features may be, to a degree, recreated by graphical conventions, the choice of reporting verbs but also body language descriptions that accompany speech. In fictional texts, for example, suspensions have been identified as "associated with specific types of body language presentation" (Mahlberg *et al*., 2020: 150). A 'suspended quotation' is defined by Lambert (1981: 6) as "protracted interruption by the narrator of a character's speech". This is, according to Lambert (ibid.: 41), a place where details on suprasegmental and prosodic features of the speech frequently occur and contribute to describing dialogue that resembles an authentic one. Mahlberg *et al*. (2020: 150) stress that suspensions "can create an impression of simultaneity" – which can otherwise be challenging to the linear nature of the text.

Body language descriptions do not occur only in connection with speech. The body language of fictional characters reveals not only how the characters behave at a specific moment but also what the characters are like more generally. The most comprehensive descriptive framework of fictional body language was developed by Korte (1997). She (1997: 3–4) conceptualises body language "as non-verbal behaviour (movement and postures, facial expressions, glances and eye contact, automatic reactions, spatial and touching behaviour) which is 'meaningful' in both natural and fictional communication". Korte points out that "[i]n the context of speech, it also plays an important role in regulating the conversation; it communicates the listener's reactions to the speaker and can either complement, replace, or contradict a spoken message" (ibid.: 27). Eye-behaviour, similarly to facial expressions, is

"extremely relevant in face-to-face interaction" (ibid.: 57). Korte (ibid.) specifically mentions three types of eye-behaviour: "gaze (one person looking at another person), mutual gaze or eye contact (two persons looking into one another's eyes), and avoiding gaze." The theory of gaze has received a great deal of critical attention, including corpus stylistic approaches to literary characterisation (Johansson and Håkansson, 2019).

Korte's (1997) body language classification framework is based on types and functions of authentic body language. For our descriptive framework, we similarly rely on authentic types and functions of bodily communication as suggested by Argyle (2010: 5) (for details see Section 5). For the descriptions of eye-behaviour the most relevant functions are expression of emotions, communication of interpersonal attitudes, and functions of supporting speech. Bodily communication varies across cultures, the greatest differences being between 'contact' and 'non-contact' cultures (e.g. Argyle, 2010: 57–61). However, the expression of emotion is similar across cultures with the main difference being the degree of expressiveness and restraint (ibid.: 66). Cultural variations include, for example, conventions about laughing and crying in public, but also linguistic categorization of emotions (ibid.: 128). The levels of gaze also vary between cultures. All cultures have norms that constrain gaze behaviour – "children are instructed to 'look at me', not to stare at strangers and not to look at certain parts of the body […] people have to look in order to be polite, but not to look at the wrong people or in the wrong place" (ibid.: 158).

Linguistically, it has been shown that in fiction, body part nouns participate frequently in multi-word combinations, whether extended units of meaning (Sinclair, 2004: 31–35; Ebeling, 2014; Mahlberg *et al*., 2020), recurrent sequences of words or collocational patterns allowing for some variation (Mahlberg, 2013). Such recurrent patterns also provide general insights into fictional characters' characterisation and their communication (Mahlberg, 2020: 144). This seems to hold across languages (Vaňková *et al*., 2005; Stubbs, 2007; Lindquist and Levin, 2008; Wieçławska, 2012; Ebeling, 2014) and applies to children's literature too. The close connection between eye-behaviour and direct speech in children's literature is supported by the fact that in our BNC children's literature sub-corpus (for details see Section 3), the most frequent 4-grams[3] containing a form of the lemma EYE are <*his eyes. '*> and <*her eyes. '*>. Both include an opening single quotation mark as their last token, indicating the beginning of a direct speech, as in *He glanced around slowly, blinking* ***his eyes. 'What happened?'*** (CFJ). These patterns are in line with Mahlberg *et al*.'s (2020: 150) observation that "[it] is typically the narrator who describes characters' body language, while accounts of body language are less frequent in the speech of characters", which is also shown by the third person possessive pronouns.

3. Data and methodology

In this study, we focus on eye-behaviour that occurs exclusively in the vicinity of speech. We do not deal with verbs of 'looking'; but we specifically explore the lemma EYE, OKO and SILMÄ respectively. We use three broadly comparable datasets of children's fiction. For this study, children's fiction is, as a text-type, defined by its audience, as texts "written to be read by children and young people" (Reynolds, 2011: 1; for limitations of this definition see, e.g., Reynolds, 2011: 1–5). In our case, the study is substantially limited by text availability. We rely on corpus compilers and their classification. For English, we use the subcomponent of

[3] The 4-grams (recurrent four-word sequences) were identified using the KonText interface; punctuation was treated as a word-token; the position of the lemma EYE in the 4-gram was not fixed.

children's fiction (2 million words)[4] in the British National Corpus (BNC), for Finnish, we use the subcomponent of original (non-translated) Finnish children's literature texts in the Savokorpus[5] (0.5 million words) and for Czech, we use a subcorpus of Czech children's books selected from the Czech National Corpus[6] (2.8 million words). For an overview see Table 1 and for the detailed composition of the subcorpora see the Appendix. Both the BNC and the Czech data were examined using the KonText interface (Machálek, 2020), the Finnish data was processed with LancsBox (Brezina *et al*., 2018).

Table 1. Corpora used in the study.

	ENGLISH	CZECH	FINNISH
Source corpus:	BNC subcorpus	Syn-7 subcorpus	Savokorpus
No. of words:	2 mil.	2.8 mil.	0.5 mil.
No. of texts:	77	59	24
No. of authors:	44 + 31 adapted classics	43 authors	19 authors
Publication dates:	1960-1994	1967-2013	1994-1999

Methodologically, Mahlberg *et al*. (2020: 144) make a strong case for a "lexically-driven approach that describes body language on the basis of repeatedly occurring linguistic patterns, in the form of repeated sequences of words". This approach has proven cross-linguistically extremely challenging between typologically different languages (for discussion see Čermáková and Chlumská, 2017; Šebestová and Malá, 2019). We, therefore, approach the analysis of eye-behaviour through the examination of the grammatical and textual functions of the nouns EYE, OKO and SILMÄ. We use lemmata because both Finnish and Czech have an extensive number of forms per noun. For each language, we first retrieved all occurrences of EYE, OKO and SILMÄ. These were further narrowed down only to occurrences of 'eyes' within the vicinity of speech, which we defined as a +/– 5-word span from the beginning or end of direct speech. For each language, we have analysed a random sample of 100 concordance lines.

The English and Czech corpus data are lemmatised. There were 2,251 total occurrences of EYE, and OKO occurred 3,986 times. In English and Czech, direct speech is marked by quotation marks, so for the selection of the relevant occurrences, i.e. those in the vicinity of direct speech, we relied on punctuation. EYE occurred near speech 372 times and OKO 1,380 times. The concordance lines were shuffled and the initial 100 instances of EYE/OKO in the narrator's speech were selected for further analysis. For Finnish, we had unlemmatised data; we, therefore, searched in LancsBox for the word root <silm.*> and relevant occurrences were selected manually. As the typographical conventions in Finnish do not use quotation marks, the relevant examples had to be extracted manually too; see example (1), where direct speech is marked in bold. After the examination of the results based on the query for the root 'silm', we identified 932 occurrences of SILMÄ. Further examination of these 932 instances for occurrences in the vicinity of speech narrowed our dataset to 168; out of this sample, we selected 100 examples aiming at even distribution across the source texts.

[4] The BNC subcorpus was defined on the basis of audience ('child/teenager'), domain ('imaginative') and medium ('book').
[5] Provided by the courtesy of Prof. Anna Mauranen.
[6] Available at www.korpus.cz. The subcorpus was defined on the basis of source language ('Czech'), audience ('children/teenagers'), text-type ('fiction'), and medium ('book').

(1) **Ai niin, käytännössä**, sanoi Joulupukki ja hänen silmissään tanssi nauru. - **Käytännössä sinä lähdet** … [sla002]
[**Well, in reality**, the Father Christmas said and there was laughter in his eyes. – **In reality, you will go** …]

The three 100-line samples were analysed from two points of view: a) the lexico-grammatical perspective (Section 4), and b) the eye-behaviour and narrative function perspective (Section 5). The lexico-grammatical analysis focuses on the relationship between the grammatical characteristics (syntactic function, case-marking, prepositions) and the lexico-grammatical patterns in which EYE, OKO and SILMÄ occur, with focus on the semantics of the co-occurring verbs.

For the analysis of the characteristics of fictional eye-behaviour we draw on authentic eye-behaviour, specifically types and functions of bodily communication as described by Argyle (2010: 5); for more details see Section 5.

4. Lexico-grammatical perspective

The lexico-grammatical perspective focuses on differences and similarities in syntactic structure in the three languages in relation to particular situations in which the eyes are involved, as indicated by the meaning of the predicate verb; here we also consider the information structure perspective.

4.1 Syntactic functions

As pointed out by Burgoon *et al*. (2010: 4–5), nonverbal behaviours, such as smiling, crying or staring in a threatening manner, "allow people to communicate with one another at the most basic level regardless of their familiarity with the prevailing verbal language system."
There are, however, differences in how the same eye-behaviour and communication are rendered by different languages, depending on their typological characteristics.

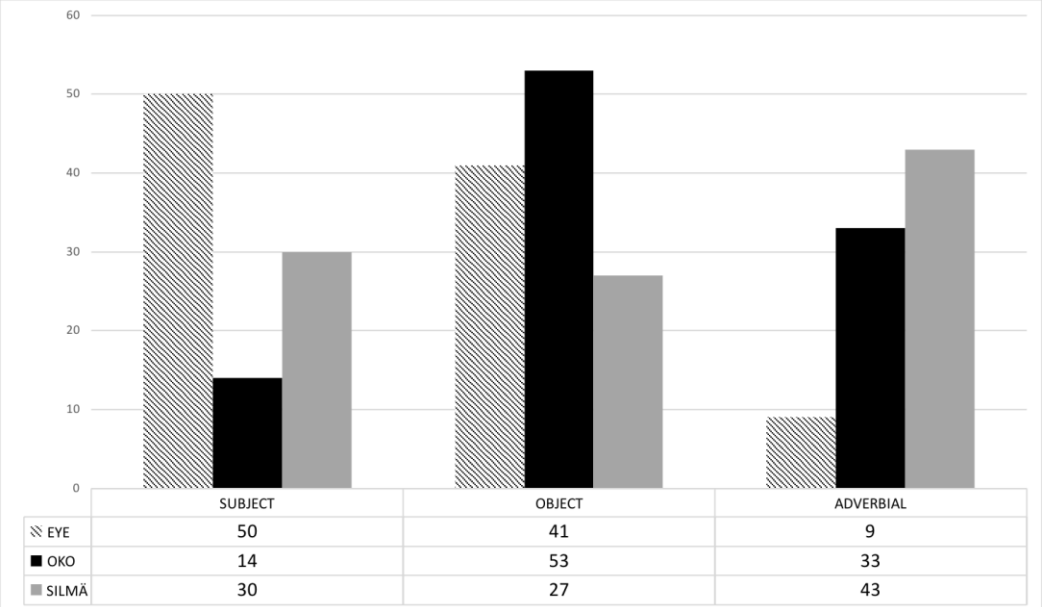

Figure 1. Syntactic functions of the 100 phrases comprising EYE, OKO and SILMÄ.

While the 100 phrases with EYE, OKO and SILMÄ perform the same syntactic functions in English, Czech and Finnish, the distribution of the functions differs (cf. Figure 1), and so does the syntactic structure of the phrases. EYE, OKO and SILMÄ typically function as the head of a noun phrase. The syntactic function of the adverbial or object may be performed by a prepositional phrase with a prepositional complement realized by an EYE/OKO noun phrase in English and in Czech (e.g. *into her eyes*). In Finnish, the object is realised by SILMÄ in nominative, accusative or partitive case. The Finnish constructions '*silmät* + E-infinitive_instructive'[7] (e.g. *silmät palaen* 'eyes shining') (example 2a) and '*silmät* + adj._essive' (e.g. *silmät suurina* 'eyes big')[8] (example 3a) were categorised as subjects analogically to corresponding structures in English (examples 2b and 3b respectively).

(2) a. Jassu kyseli **silmät palaen** … [sla001]
[Jassu was inquiring eyes E-INF.INSTRUCTIVE.-shining …]

b. … the doctor said, **eyes twinkling** [FSR]

(3) a. Anisa tuijotti **silmät suurina** hänen olkapäänsä yli. [sla010]
[Anisa stared eyes ESS.-big across her shoulder]

b. Odhar continued, **his eye hard** on his son [APW]

In the 'subject' category, we have also included Czech and Finnish noun phrases where OKO or SILMÄ is a postmodifier/premodifier in the genitive case (e.g. *tři páry překvapených* ***očí*** – 'three pairs of GEN.-surprised eyes', ***silmien*** *katse* – 'look of GEN.-eyes'). No such subject phrases occurred in our English data.

There is a statistically significant difference between the number of EYE, OKO and SILMÄ phrases used as the subject in English (50 occurrences), Czech (14) and Finnish (30) respectively. There is no significant difference between adverbial uses in Czech and Finnish (33 and 43 cases respectively); English (9) differs significantly from both. While there is no significant difference in the object use in English and Czech (41 and 53), Finnish (27) differs significantly from both English and Czech.[9]

These differences can be ascribed to typological distinctions. English and Czech

> involve different hierarchies of the operating word order principles: owing to its analytic character, English employs word order primarily to indicate grammatical functions; on the other hand in inflectional Czech the grammatical principle plays a secondary role, syntactic relations being indicated by grammatical endings. Hence Czech word order is free to perform other functions among which indication of the FSP ['functional sentence perspective', i.e. information structure] functions of the clause elements ranks highest. (Dušková, 2015: 14)

Finnish is closer to Czech with respect to the role of inflection and word order: it "exhibits relatively few constraints on word order in a finite clause […], case suffixes guide arguments into their canonical positions […], the word order correlates with discourse interpretation" (Brattico, 2020: 38–39). Word order in Finnish finite sentences is "constrained by information structure": there are "designated word order positions for the topic of the sentence and a phrase that carries a contrastive focus" (Nikanne, 2017: 69).

[7] E-infinitive is also referred to as II. infinitive.
[8] Hakulinen (ed.) (2004: 837 § 877) calls these 'status constructions', which have an adverbial function expressing state of being.
[9] The difference is statistically significant at the 0.05 level of significance (Log-likelihood). Cvrček, V. (2021). *Calc: Corpus Calculator* 1.02. Prague: Czech National Corpus (Available from https://www.korpus.cz/calc/) was used to calculate statistical significance.

4.2 The subject and the adverbial

The most marked differences in the distribution of syntactic functions in the three languages are between the subject and adverbial. Half of the instances of EYE-phrases in our English sample function as the subject, occupying the initial position in the sentence. The 'eyes' are thematic elements, carrying a relatively low amount of information; their contextual dependence is signalled by anaphoric possessive determiners or definite articles, see example (4a). The predicate verb is typically intransitive or copular, with the absence of further complementation making it the most prominent, rhematic part of the message. Semantically, the EYE-subjects display a preference for predicates which express the emission of light, such as SHINE, BE BRIGHT or FLASH (see also Section 5.1.1 for discussion of light metaphor), movement or absence thereof: BE DRAWN, BE FIXED, BORE, CLOSE, NOT LEAVE, REACT, REST, RISE, ROLL, SEARCH, SLIDE, STAY, TURN, WANDER, and a change, usually in the shape of the eyes: BULGE, NARROW, ROUND, SHARPEN, SOFTEN.

The semantically corresponding situations may be rendered in Czech as clauses with a divergent syntactic structure. In clauses drawing on the light metaphor, the OKO-phrase is often constructed as an adverbial in Czech (example 4b). Although the syntactic structure is different in examples (4a) and (4b), the information structure is similar in the two languages. The clause-initial position is occupied by a thematic, context bound element, expressing the location of the 'light'; the phrase functions syntactically as the subject (*his eyes*) in English and as the adverbial (*v očích* 'in eyes') in Czech. The locative semantics of the OKO-phrase is supported by the preposition *v* ('in') and case marking (the locative case) in Czech. In both languages, the predicate verbs are intransitive[10] and the predicates constitute the focus of the message.

In Finnish, in about a third of the occurrences where verbs of 'light' occur, the situation is syntactically very similar to Czech, and in terms of information structure it is similar to both Czech and English. 'Eyes' occur in sentence initial position in one of the 'locative' Finnish cases (inessive, illative, elative);[11] see example (4c). In most of the remaining cases, Finnish is syntactically closer to English, 'eyes' occur as thematic in the subject position. For this semantic group of verbs, the typical construction in Finnish seems to be the E-infinitive in instructive case as in example (2a) above. (For a discussion of infinitives in Finnish see, for example, Toivonen, 1995; Hakulinen (ed.), 2004.)

(4) a. 'Great!' cried Mould. **His eyes** shone brighter … [AMB]

b. **V očích** jí najednou blýsklo. [Bílá ruka a poklad hradu Handštejna]
[In LOC.-eyes to DAT.-her suddenly flashed.]

c. **Silmistä** paistoi uteliaisuus … [sla006]
[From ELAT.-eyes shone curiosity …]

The clauses with intransitive verbs indicating movement or absence thereof, with the EYE-phrase as its subject in English (example 5a), correspond semantically to Czech clauses that either display the same syntactic structure or – more frequently (11 instances) – render the OKO-phrase as an adverbial (example 5b). The adverbial is formed by a noun phrase with OKO in the instrumental case indicating the means or instrument of 'looking'.[12] There were no examples of intransitive verbs of movement attested in the Finnish sample. However, there were several cases corresponding to the instrumental use. These occurred together with verbs of 'looking'

[10] In Czech the clause in example (4b) is subjectless; there is, therefore, no "competitor" of the verb in terms of information load (cf. Firbas, 1992: 7).
[11] Inessive having a locative meaning "inside", illative meaning "into" and elative meaning "from".
[12] The eyes were constructed as instrument (a *with*-prepositional phrase) in one clause only in the English data: *Mortimer scrutinized her **with narrowed eyes** …* [FSR]

with SILMÄ in the adessive case; this is thus more similar to English both in terms of syntactic and information structure. See example (5c) for Finnish and footnote 11 for English.

(5) a. … **her eyes** wandered doubtfully to Ferryman [AEB]

b. "Má bejt …?" utrhl se, a **roztěkanýma očima** bloudil po stadiónu. [Metráček]
["So what …?" he snapped, and with INSTR.-restless eyes wandered around the stadium."]

c. … ja vieras tyttö katseli sivusta **sameilla silmillään**. [sla004]
[…and the foreign girl look sideways with her ADESS.-cloudy eyes.]

There were no Czech clauses in the sample similar to the English occurrences of intransitive verbs indicating change in the shape of the eyes (e.g. *rounded* in example 6a). In Finnish, we found two examples, both with the verb *pyöristyä* ('round') (example 6b). Both the syntactic and information structure are divergent here, though.

(6) a. **Nick's eyes** rounded with remembered horror. [EFJ]

b. Tämä sai **Nannan silmät** pyöristymään. [sla020]
[This got **Nanna's eyes** to MA-INF.ILLAT.-round.][13]

Apart from the correspondences between the English subject and the Czech adverbial as explained above, the high number of OKO-phrases in adverbial position can be accounted for by occurrences where the direction of the gaze is specified by a prepositional phrase with the preposition '*do* ('into') + GEN.-OKO' (example 7a, 11 instances). Four corresponding constructions can be found in the English data (example 7b). Finnish expresses both directions: 'into' (illative case, 18 occurrences) and 'from' (elative case, 3 occurrences). The 'into' direction, similarly as in English and Czech, accounts for occurrences of 'looking into sb.'s eyes' (example 8). The remaining 10 illatives account for cases where something else is happening to the eyes; these typically occur with verbs of an upward movement, as NOUSTA ('rise') in example (8) (for further discussion see Section 5.1.1).

(7) a. Hluboce se nadechla, pohlédla matce **do očí** a otázala se … [Vládci sedmihoří]
[She took a deep breath, looked her mother **in** GEN.-**the eyes** and asked …]

b. I looked deeply **into her eyes**. [FRU]

(8) **Anisan silmiin** nousivat kyyneleet. [sla010]
[Into Anisa's ILLAT.-eyes rose tears.]

There are no corresponding examples in English and Czech for the Finnish elative ('from') constructions illustrated in example (9).

(9) Kyyneleet valuivat **peikkomuorin silmistä** ja tipahtelivat sileälle kivelle. [sla022]
[Tears were running from the old troll woman ELAT.-eyes and were falling on a smooth stone]

As discussed above, most of the adverbial examples in Finnish are accounted for by very precise encoding of location that is inherent to the language system. In addition to the cases discussed above, inessive case (meaning 'in/inside') is represented in our sample by 12 occurrences.

Another factor contributing to the preponderance of the adverbial function of the OKO-phrases in Czech and of the subject EYE-phrases in English is related to the difference in the representation of non-finite and verbless supplementive clauses in the two languages. English uses these clauses for backgrounding information on the accompanying circumstances and

[13] MA-infinitive is also referred to as III. infinitive

simultaneous actions[14] while contributing to the impression of the authenticity of the speech (example 10a, with EYE as the subject of a participial clause). In Czech, verbless and non-finite clauses are relatively rare (cf. Malá and Šaldová, 2015). The corresponding adverbial meanings can be expressed by a prepositional phrase (example 10b); simultaneity can be inferred from the coordinative relation between two finite clauses linked by the coordinator *a* ('and') or juxtaposed (example 24 below). The sequential or simultaneous interpretation of the action performed by the eyes and speech is also supported by the verbal aspect in Czech, with the perfective aspect expressing completed, bounded actions (cf. the verb forms *pohlédla* 'looked' and *otázala se* 'asked' in example 7a), and the imperfective aspect actions or processes in progress (cf. *dívá se* 'is looking' in example 31) (Cvrček *et al*., 2015: 292). In Finnish, the impression of simultaneity is typically expressed with the E-infinitive in instructive case (6 occurrences) and adjective in essive (5 occurrences), see examples (2a) and (3a).

(10) a. '… I think, though,' the Doctor said, **eyes** twinkling, 'we'll be able to persuade them …' [FSR]

b. "Kde ty se tu bereš?" řeknu **s vytřeštěnýma očima** klukovi stojícímu přede mnou. [Nová láska na obzoru]
["What are you doing here?" I-say **with** INSTR.-**wide-open eyes** to DAT.-the boy standing in front of me.]

4.3 The object

The difference in frequency of the object function of the EYE- and OKO-phrases between English and Czech is not significant. In both languages, the eyes function most frequently as the object of verbs of 'opening' or 'closing', with the variety of verbs constituting a scale between eyes wide open and closed being broader in Czech: VYTŘEŠTIT, VYPOULIT, VYVALIT, (VY)KULIT, OTEVŘÍT, POOTEVŘÍT, PŘIMHOUŘIT, ZAMHOUŘIT, PŘIVŘÍT, ZAVŘÍT. All these verbs are in the perfective aspect that here conveys a change of state of the eyes. The prefixes *po-* and *při-* make it possible to express a lesser degree of 'openness'. Many of these verbs are emotionally coloured. The corresponding English scale comprises merely OPEN, SHUT, CLOSE, KEEP CLOSED. SILMÄ-phrases function as object significantly less frequently in comparison with Czech and English. However, similarly to English and Czech, they are frequently attested with verbs of eye 'opening' (AVATA) and 'closing' (SULKEA).

Another frequent semantic class of verbs which take EYE- and OKO-phrases as their object includes verbs indicating the direction in which the eyes move: DROP, TURN (ON), FIX (ON), KEEP (OFF/ON), REFOCUS, ROLL in English; SKLOPIT, ZDVIHNOUT, ZVEDNOUT, PŘEVRÁTIT / ZVRÁTIT V SLOUP, NEODVRACET, ODLEPIT, OBRÁTIT, POZVEDNOUT, PŘIŠPENDLIT in Czech. The eyes also occur as the object in descriptions which do not directly relate to the adjacent speech. While in the Czech data this is less frequent with only one verb falling in this category (UTÍRAT 'wipe'), in English the range of verbs is broader: WIPE, SHADE, SHADOW, RUB, COVER. The SILMÄ-phrase, similar to English, also frequently occurs in body language descriptions that do not directly relate to speech: eyes are being 'dried' (KUIVATA), 'protected' (SUOJATA), 'covered' (PEITTÄÄ) and 'rubbed' (HIEROA). The SILMÄ-phrase as object also occurs in possessive constructions 'adessive + *olla*'.[15]

[14] There were seven non-finite and verbless clauses with the EYE-phrase as the subject in the English sample. In the two Czech verbless clauses in our sample, OKO functioned syntactically as the object.

[15] There is no verb 'have' in Finnish. The 'possessor' is in the adessive case followed by the verb 'be' and the 'possessed' is frequently interpreted as 'object', e.g. …*hänella oli tummat kulmakarvat ja* ***silmät*** ('ADESS.-s/he had dark eyebrows and **eyes**') [sla010]. However, Hakulinen (2004: § 895–898), for example, classifies these as a subtype of existential constructions.

Our samples are too small for us to describe the lexical patterns with EYE, OKO and SILMÄ systematically; however, based on these limited samples, they seem to frequently participate in phraseological/idiomatic constructions. In Czech, the eyes are used as the object of a number of verbs which are severely restricted in their collocability with other nouns as objects, such as (PŘI/ZA)MHOUŘIT ('narrow, squint'), (VY)TŘEŠTIT ('open wide'), (VY)KULIT or (VY)POULIT ('pop'), PROTŘÍT ('rub'), and in idiomatic expressions, for instance OBRÁTIT/PŘEVRÁTIT/ZVRÁTIT *oči v sloup* ('roll one's eyes upwards'), LHÁT/ZALHAT *do očí* ('lie, pull the wool over somebody's eyes'). Other idiomatic expressions include 'hodit okem po', which is similar to the Finnish 'iskeä silmä' ('throw an eye on'). The idiom 'believe one's eyes'/ 'věřit svým očím' / 'uskoa silmiään' seems to be common to all three languages.

There were fewer verbs with restricted collocability in English and Finnish; these include for example, STARE, ROLL, BLINK or SIRISTÄÄ ('narrow, squint') and TUIJOTTAA ('stare'). The phrase 'corner of her/his eye' occurred repeatedly and it did not have a semantic or functional equivalent in our data in Czech or Finnish. The English phrase 'keep an eye on' had its Finnish equivalent 'pitää silmällä'.

5. Fictional eyes: characteristics and narrative functions

The linguistic description of eye-behaviour accompanying fictional speech shows how characters behave before, during or after speaking. Similarly, as the functions of fictional speech reflect the speech in the "real" world to the extent that it supports the characterisation and the narrative (Leech and Short, 2008; Semino and Short, 2004; Mahlberg *et al*., 2019), we can expect the degree of faithfulness to "real" eye-behaviour to be on a cline. In Section 5.1, we will focus on the characteristics of fictional eye-behaviour and in Section 5.2 we offer another perspective: we will aim to capture the narrative functions, that is, why the reader's attention is drawn to the characters' eye-behaviour in the first place. The delimitation of the narrative functions is, as can be expected, not straightforward. Unlike in language, there are no standards of form (Beattie, 2004: 79). Despite this variety, we have, as expected, observed similar characteristics and narrative functions across all three languages.

5.1 Characteristics of fictional eye-behaviour

Argyle (2010: 5) defines five types and functions of bodily communication: (i) expressing emotions, (ii) communicating interpersonal attitudes, (iii) accompanying and supporting speech, (iv) self-presentation and (v) rituals. Eye-behaviour plays a central role in the first three (Sections 5.1.1 to 5.1.3). Central to eye-behaviour is looking, or 'gaze'. Looking is primarily "a means of perceiving the expressions of others" but "the act and manner of looking also have meaning as signals, showing for example the amount of interest in another person… So gaze is both signal and channel, a signal for the recipient, a channel for the gazer" (Argyle, 2010: 153). There are different aspects of gaze that have been considered: the amount of gaze at other, mutual gaze, looking while talking and while listening, pupil dilation, eye expression, direction of gaze-breaking, or blink-rate (Argyle, 2010: 153–154).

All these aspects are more or less frequently present in fictional texts. Korte (1997: 58) considers the direction and duration of the gaze as the most determining expressive quality. Explicitly expressed mutual gaze and direct eye contact are less frequent than perhaps expected. In our sample, it occurs only eight times in English, nine times in Czech and eight times in Finnish. In Czech, all these occurrences build on verbs of 'looking' – in Czech: 'DÍVAT

SE/PODÍVAT SE/POHLÉDNOUT + *do očí*' ('look into eyes').[16] In Finnish, the most frequent construction is 'KATSOA/TUIJOTTAA *silmiin*' ('look/stare into sb.'s eyes'). In English, in addition to 'LOOK/STARE + *in/into* sb.'s *eyes*', we have also found examples where mutual gaze is described in other ways, see example (11). A similar example was found for Finnish as well (example 12).

(11) His **eyes** were still **fixed on mine**. [FPU]

(12) mä vastasin ja yritin **pitää silmäni sen silmissä**. Entä muuten? se kysyi [sla019.txt]
[I answered and tried to keep my eyes in his INES.-eyes. So what else? he asked…]

Occurrences where one person is looking at another are more frequent than explicitly expressed mutual gaze in all three languages and there is also a greater lexical variation. In English, for example, eyes TURN *to*, LOOK (*up/at*), REST/STAY *on*, REFOCUS *on*, SCRUTINIZE or simply *are* (*fixed*) *on*, see example (13), with the target of 'looking' being typically the person or their face.

(13) "And then," said the boy, **his eyes on Doyle's face**, "then you'll shoot me." [AC4]

In Czech, examples include 'HODIT *okem po*' ('throw an eye on'), 'OBRÁTIT *oči po*' ('turn eyes to'), '*oči probodávají*' ('eyes drill'), '*oči se přibližují k*' ('eyes are coming close to'). In Finnish, we find 'ISKEÄ *silmä*' ('throw an eye'), 'KATSOA' ('look'), '*silmät porautuu*' ('eyes drill'), '*silmät tutkailee*' ('eyes scrutinize').

Avoiding gaze, or gaze breaking, is relatively infrequent in all three languages. In Czech, the specific verb SKLOPIT ('cast down') occurs three times (in two source texts), see example (14).

(14) „Tak promiň," **sklopím oči** a tvářím se nešťastně a ukřivděně. [Když přijde láska]
["Sorry," I cast down my eyes and look unhappy and aggrieved.]

We do not find corresponding examples in English and Finnish. Several occurrences of gaze breaking in English exemplify situations where the character is showing lack of interest, e.g. (15). There was also an example when a character breaks eye contact in order to make eye contact with someone else, e.g. (16).

(15) "Anybody could walk in." **Bella's eyes were fixed on** the television screen: she didn't even turn her head. [ACB]

(16) "Oh no, she won't be angry," Nick said. **His eyes slid slyly sideways** at Carrie and he started to giggle. [EFJ]

While direction, duration and intensity of the gaze are frequently lexically expressed through the choice of verbs, adverbial constructions and prepositional phrases (see Section 4.1), from the reader's point of perspective, it is often a complex decoding process, as shown in example (17).

(17) Mungo was about to say "yes", when there was a bellow from the direction of the pub. "SHOP!" Lily **rolled her eyes**. "You see what I'm up against?" she appealed. "Pig ignorant they are." As she **turned** to serve the impatient customer she added: "I've been in palaces and kings' houses, Mr Stone. […]" [ACV]

In this example, Lily 'rolls her eyes' expressing both emotion and attitude. Though not explicitly mentioned, Lily is probably also looking at the person she is speaking to, as she is then 'turning away' to serve her customer. Again, though not directly mentioned, we can

[16] All these verbs have very similar meanings: DÍVAT SE is a reflexive verb in imperfective aspect meaning 'look', PODÍVAT SE is in the perfective aspect, it is based on the same verb with the prefix *po-*, which suggests the "looking" is brief and quick, and the perfective POHLÉDNOUT is stylistically slightly archaic and also refers to a brief and quick look.

assume she is also being looked at while speaking. So, while textually only one aspect of the eye-behaviour is described, *rolled her eyes*, the reader, drawing on their experience of similar situations – both fictional and "real-world", will interpret the text more holistically.

5.1.1 *Expressing emotions*

In addition to eyes, emotions are primarily expressed in the face but also in the body and voice. Emotions may be expressed spontaneously, attempted to be controlled in order to conform to social rules, or concealed for other reasons (Argyle, 2010: 4). Emotions are "classified in terms of dimensions: the dimensions most commonly found are pleasant-unpleasant, and level of arousal" (ibid.: 72). Emotions in the eyes are primarily conveyed through the amount of eye opening, pupil dilation and amount of gaze (ibid.: 73). Facial expressions of emotion are cross-culturally similar and, indeed, our data show similarities across the languages rather than differences. However, the linguistic repertoire is extremely rich as we have shown in the previous section and there are clear differences in syntactic preferences between the languages that are also reflected on the semantic level.

Emotions, both pleasant and unpleasant, may be described explicitly and directly in all three languages, as exemplified in example (18).

(18) "No, look –" The big blue **eyes were full of pain, innocent, apologetic**. "I want to help you, honest … [AEB]

In example (18), the eyes are described with pre-modifying adjectives in terms of their size and colour[17] and the emotions are described as if independently as a copula complement. Alternatively, it can be the premodifiers of eyes that convey the emotional states, e.g. *steady*, *icy*, or *sharp* in English; PŘEKVAPENÝ ('surprised'), VYDĚŠENÝ ('scared') and ROZTĚKANÝ ('distracted') in Czech (see example 5a above); and VÄSYMYKSEN TÄYTTÄMÄ ('full of tiredness'), TYHJÄ ('empty') and TOTINEN ('serious') in Finnish. In other cases, eye-behaviour is also described, as in example (6a) above, where Nick's eyes' expression changes, in that his eyes round and the reason for "rounding" is horror, which is similar to example (19) below, where eyes are being nearly closed also with horror.

(19) Tu prosí potřetí: „Bělinko, ženo drahá, polib mě!" **S hrůzou zamhouřila** Běla **oči** a políbila hada. [Sedmero krkavců a jiné pohádky]
[And he pleads for the third time: Belinka, my dear wife, kiss me!" Bela **closed her eyes with horror** and kissed the snake.]

In other cases, we find emotions described directly in the speech and the description of the eye-behaviour amplifies the content. In example (20), the intensity of the moment is communicated by describing the directed gaze.

(20) The monster's **yellow eyes looked at me**. "I am the **unhappiest** creature in the world, but I shall fight for my life," he said. [H8G]

Often, however, it is less straightforward to decode the emotion, as in example (21). For the intended reading of this example, the reader needs to be familiar with the fact that "round eyes" (*silmät pyöreinä*) signal surprise.[18] For the child reader, contextual situational cues, as in

[17] The colour of the eyes seems to be of particular importance in Finnish, with colour terms constituting 60 per cent of the modifiers of eyes, e.g. MUSTA, HARMAA, KELTAINEN, UTUISENVIHREÄ, HAALEA or KAISLANVÄRINEN ('black', 'grey', 'yellow', 'hazy green', 'pale', 'reed-coloured'). In Czech the colour and size modifiers of the eyes are rare.

[18] Though not occurring in our sample, 'round eyes' signal surprise in English as well: "She stopped and her **eyes grew round with surprise**." [BOB]. In Czech, the 'roundness' indicating surprise is encoded primarily through the verb VYKULIT, which has the stem 'kul-', on which words signifying 'round' are based, e.g. adj. *kulatý* 'round'.

example (21) where Rietta unexpectedly appears on the scene, also help them arrive at the intended reading.

(21) Mitä sinä täällä teet? Vaaputin katsoi **silmät pyöreinä** Riettaa. [sla021]
[What are you doing here? Vaaputin **was looking** at Rietta **with round eyes**]

One of the characteristic features of fictional eye-behaviour that occurs frequently across all three languages is the expression of emotion through a light metaphor (see also Section 4.1.1). While pupil dilation is rarely explicitly mentioned in literary texts (Korte, 1997: 58), the light metaphor can be interpreted in relation to the amount of eye opening, expression but also pupil dilation. Light metaphor occurs both with pleasant and unpleasant emotions and can occur in either the speaker's eyes to support the speech, or the listener's eyes to manifest the reaction to what has been said. In the majority of the occurrences, the eyes go from a darker to a lighter state.

In English, the verbs encoding the light metaphor in our sample are SHINE, BURN, FLASH, GLOW, GLEAM, TWINKLE, LIGHT; eyes are also repeatedly described as *bright*. In Finnish, the verbs are LOISTAA ('shine'), VÄLÄHTÄÄ ('flash'), KIILUA ('glow'), LEIMUTA ('flame'), PAISTAA ('shine'), PALAA ('burn'), PILKAHTAA ('twinkle'), SÄTEILLÄ ('radiate'), SYTTYÄ ('ignite') and TUMMUA ('darken'), the rare case where the metaphor is expressed from light to dark. The noun *valo* ('light') also occurs (*silmissään käy outo valo*, 'a strange light appears in the eyes'). In Czech, the verbs include (ZA)SVÍTIT ('shine'), ZALESKNOUT ('shine'), ZATŘPYTIT ('glitter'), ZAJISKŘIT ('sparkle'), (ZA)PLÁT ('burn'), BLÝSKAT ('flash'), POBLÝSKÁVAT ('flash'), and also ZATMĚT ('darken')[19], see examples (22) to (23).

(22) "What do you want?" "Me? Experience, mostly. I want to know things," said Gay. **Her eyes**, which were very blue, **burned for a moment like sapphire lamps**. [BMU]

(23) Liito rypisti otsaansa, ja hänen kaislanväriset **silmänsä tummuivat harmista**. – Tietysti kiusasivat, hän tuhahti. [sla002]
[Liito wrinkled his forehead and **his** reed-coloured **eyes darkened with annoyance**. – Of course, they bullied, he sniffed.]

(24) „Vašku, nech toho!" okřikla ho rozzlobeně babička a **oči jí hněvivě zaplály**. „Pan Havránek nám pouze chce pomoct. [Klobouky z Agarveny]
["Vasek, leave it!" nan shouted at him angrily and **her eyes were glowing anger**.]

Emotions are also described through other typical accompanying emotional signals: tears or crying[20] are frequent, but laughter also occurs. In relation to 'tears' and the direction of eye-movement, there are some differences between the languages. While in English 'eyes are full of tears', 'tears fall from eyes', 'eyes fill with tears'; in Czech "slzy vstoupí do očí" ('tears enter into the eyes'), "slzičky se zatřpytí v očích" ('little tears shine in the eyes'), someone 'has tears in the eyes' ('MÍT slzy v očích'); in Finnish tears 'fill eyes' ('silmät täyttyy kyynelistä'), 'fall from eyes' ('kyyneleet valuu silmistä') or someone 'has tears in the eyes' (*Petellä oli tosiaan kyyneleet silmissä*, 'Pete really had tears in his eyes') but typically 'tears' 'rise into someone's eyes', see example (25). Both in Finnish and in Czech, therefore, tears can play a more active role, moving into the eyes.

(25) Hyvää matkaa, hän sanoi ja **hänen silmiinsä nousivat kyynelet**. [sla010]
[Have a good trip, he said and **tears rose to his eyes**]

The verb NOUSTA ('rise') in Finnish also occurs with *ilme* ('expression') and *katse* ('look') that rises into someone's eyes (in illative case which is the case expressing direction) (see the

[19] Many of these verbs occur with prefix *za-*, which signifies short duration and stresses the beginning of the action.
[20] These can be also described less directly as 'wiping her eyes' or 'kosteat silmät' [moist eyes].

discussion in Section 4.1.1). This is different from both English and Czech. In English, the subject of the verbs that have the semantic feature of an upward movement is 'eyes', as in example (26), and in Czech it is the person who moves the eyes upward (ZVEDNOUT, ZDVIHNOUT), as in (27). The upward movement into the eyes appears to contribute to the expression of emotion in Finnish, while in English and Czech the movement of the eyes primarily accompanies and supports speech.

(26) Her huge **eyes**, gleaming hazel, **rose to his**, triumph carefully hidden. [APW]

(27) Kája maličko zaváhal, než **zdvihl oči**: „Já bych, prosím, tuze rád, ale naše maminka by asi nechtěla. [Školák Kája Mařík]
[Kaja hesitated a bit before he lifted his eyes: "I would, very much like to, but my Mum would probably not want me to."]

5.1.2 *Communicating interpersonal attitudes*

Interpersonal attitudes and relationships are primarily communicated through physical proximity, tone of voice, touch, gaze and facial expressions (Argyle, 2010: 5). In many respects, attitudes are very similar to emotions and may involve exactly the same signals (ibid.: 86). Gaze can communicate, for example, liking through intensity and duration and through mutual gaze (ibid.: 88). Another type of attitude that is being established through gaze is dominance. In an established hierarchy, less gaze but more looking while talking signals dominance, while more gaze and staring other down signal attempts to actively establish dominance (ibid.: 97); see examples (28) and (29). In example (28), Miss Jarman signals her dominance and anger, her gaze is intensive and intimidating – *her sharp eyes bored like drills*. In example (29), the angry gaze is a response to what has been said. We know that the *glowing eyes* of the listener have been decoded by the speaker as anger and the speaker responds with *a raised warning finger*.

(28) "What's that? Speak up. Raise your head. Climbed up what if you please?" "The mooring rope." Miss Jarman's **sharp eyes bored like drills**. "Because?" [C85]

(29) "He's finished in Dresden and he's coming back tomorrow." Omi's **eyes glowed** but Frau Nordern **raised a warning finger**. [A7A]

Attempts at establishing dominance can be described through expressions of the intensity of the gaze, which is interpreted in conjunction with the speech, as in example (30), where the intensity of the look is described as 'firm' (*tiukka*), which is supported by the threatening nature of the speech.

(30) … mä sanoin hiljaa ja **katsoin sitä tiukasti silmiin**. Sä heräät yks aamu ilman korvia…[sla017]
[I said quietly and **looked him firmly in the eyes**. You will wake up one morning without ears…]

Example (31) shows a hierarchical relationship in which the person lower on the hierarchy is looking 'with innocent eyes' while example (32) shows gaze aversion as an expression of slight embarrassment.

(31) Kája **se dívá bezelstnýma očima do očí** profesorových. [Školák Kája Mařík]
[Kája is looking with his innocent eyes into the professor's eyes.]

(32) „Nelichoť mi," **sklopí oči**. [Tenhle kluk je můj!]
["Don't flatter me," she lowers her eyes.]

5.1.3 *Accompanying and supporting speech*

Argyle characterizes the 'accompanying and supporting speech' function as speakers and listeners engaging "in a complex sequence of head-nods, glances, and non-verbal vocalizations which are closely synchronised with speech and play an essential part in conversation (Argyle, 2010: 5). This is an area widely researched in linguistics. How this translates in descriptions of fictional speech is a less studied area. Our focus is on eye-behaviour that supports the fictional speech in ways that make it more "authentic", see example (33), where the character's eyes open wide, and the speech is described as high-pitch through the reporting verb *vyjekla* 'shrieked' modified by the adverb *zděšeně* ('horrified').

(33) Posléze **vytřeštila oči** a zděšeně vyjekla: „Jsme tady uvězněni, všude kolem jsou bažiny a zase bažiny." [Vládci Sedmihoří]
[After that she **opened her eyes wide** and gave a horrified shriek: "We are imprisoned here, there is nothing but swamp around here."]

Eye-behaviour descriptions that directly support the speech may also function instead of a reporting verb (see examples 14, 18, 21, 28 above). However, eye-behaviour descriptions may not only support the speech, but also function as a response to what has been said; see example (29) above.

Argyle (2010: 109) notes that "the main reason that speakers look at listeners is to obtain information, especially to obtain reactions to what has just been said". In fictional texts, this may be subtle as in example (34), where speaker's eyes remain fixed on the addressee during the speech and after the speaker has finished talking, inviting the addressee to respond.

(34) Monks listened with close attention, biting his lip and staring at the floor. "Before your father went to receive that money, he came to see me," continued Mr Brownlow slowly, **his eyes fixed on Monks' face**. "I never heard that before," said Monks, looking up suddenly, a suspicious expression on his face. [FRK]

The addressee may not, however, always respond in the expected or desired way, as illustrated in example (35).

(35) "I must apologize – I see you know the lady personally." But **he had dropped his eyes** and lost interest in me. [HGS]

5.2 Narrative functions

We have identified three broad functions related to eye-behaviour: a) eye descriptions reveal characters' attitudes and contribute to their characterisation, thus creating a relationship with the reader; b) descriptions of eye-behaviour contribute to the management of the narrative and plot creation, in that they move the narrative forward (this is specifically connected with the verbs of eye 'opening' and 'closing'); and c) eye-behaviour descriptions contribute to the overall development of the plot and description of the situation, rather than directly relating to speech. However, these functions are not easily delimited, as the functions may overlap and combine.

Our first narrative function of conveying attitudes and contributing to characterisation largely overlaps with Argyle's functions of expressing emotions and interpersonal attitudes (see Section 5.1.1 and 5.1.2). However, the function of supporting and accompanying speech (Section 5.1.3) is also largely relevant for characterisation. As discussed above, eye-behaviour may be accompanied by verbs of speaking introducing direct speech or serve as an introductory signal itself. This is often the case in Czech, a language with a high degree of lexical variation in reporting verbs (Nádvorníková, 2020), where verbs from semantic domains other than

'speaking' can introduce direct speech, see example (36). The omission of the reporting verb is also common in Finnish.

(36) Vivian *se* **zaleskly posměšně oči**: „A vešel jste a zeptal jste se na holčičku se sáňkami a ...!" [Pan Tau a tisíc zázraků]
[Vivian's eyes glistened derisively: "And you went in and asked about a girl with a sledge and…!"]

The function of narrative management and turn-taking has also been discussed already (see Section 5.1.3). As Argyle (2010: 161) notes "[g]aze plays an important role in negotiating when social encounters will start and when they end." This is similar in fictional worlds; looking and movements of the eyes and the lack thereof can be used to signal interaction among participants in conversation, indicating turn-taking (Hoffmannová, 1999: 85), or the pace of the verbal exchange. The most frequent and relevant patterns we have identified include 'LOOK *(up) at*', 'OPEN *one's eyes*', 'CLOSE/SHUT *one's eyes*' and 'FIX *one's eyes on*' (the Czech and Finnish patterns correspond, for a large part, to the English ones). Where opening the eyes or a gaze directed at the addressee precede the character's direct speech, the eye behaviour appears to be a signal so clear that it generally does not need to be accompanied by a verb of speaking in either language. The reporting verb either follows the direct speech or is missing.

A different narrative-organizing function seems to be associated with eye movement that interrupts the character's direct speech: the pace of the dialogue is slowed down, there is a pause, often described explicitly (see example 37).

(37) "...They had helmets of silver and spears like flames. Ah!" Ilbrec **closed his eyes momentarily.** "Many battles have I fought, but it is the memory of that one chills me most…" [F99]

Sometimes, the eyes provide the only reaction (example 38) in the communication, substituting for the participant's verbal response.

(38) "Adam," she said. **His eyes reacted, coming to meet hers**. He remembers who he is, Ruth thought with a pang of relief. [F99]

An action of closing and opening eyes specifically seems to have also a function of finishing a particular scene in the narrative (e.g. eyes closing in example 39), or moving the narrative forward when eyes open, as in (40).

(39) "Yes, father, I will." The King **closed his eyes** and did not speak again. [GV9]

(40) Neljännen päivän aamuna isä viimein **avasi silmänsä**. – Pikkuruu Mustanmusta, totisesti, sinä se olet… [sla022]
[On the morning of the fourth day dad finally opened his eyes. – "Pikkuruu Mustanmusta, indeed, it is you…]

Finally, the eyes may contribute to the development of the plot and description of the situation, rather than directly relating to the speech, see example (41).

(41) Joe's **eyes rolled around the room**, noticing the expensive furniture I had bought recently. [FPU]

Eyes may also be part of other body language descriptions than the categories we have discussed in the previous sections. Although these descriptions do not directly relate to the speech, they are important not only for the context of the situation but also for characterisation; see example (42), where the character blows her curls from her eyes before speaking, a gesture with a less straightforward interpretation, possibly enabling eye contact before the speech. In (43), the gesture will be more familiar, supported by the choice of the reporting verb and the following speech itself.

(42) Marke kääntyi terävästi. Toiseen poskeen painui kuitenkin hymykuoppa, hän **puhalsi kiharat silmiltään**. – Mitä? [sla004]
[Marke sharply turned. But in her second cheek a smile dimple appeared, she blew her curls from her eyes. – What?]

(43) Äiti painoi **kädet silmiensä eteen** ja huokasi: – Voi voi, mitähän tästä oikein seuraa? [sla006]
[Mother pressed her hands in front of her eyes and sighed: Oh no, what will follow from this?]

6. Conclusions

In this article we aimed to examine possible meaning-making processes of novice readers, that is children, from a cross-linguistic perspective: English, Czech and Finnish. It has been repeatedly stressed that reading is important for the development of cognitive capacities and evidence shows that stories, in particular, may contribute to enhanced social cognition (Mar, 2018), and therefore fiction texts are thought to be particularly effective in engaging young readers in meaning-making processes (Oakhill *et al*., 2015, Jerrim and Moss, 2019). Jerrim and Moss (2019: 182) hypothesise, based on previous research, that

> the cognitive demands that extended narrative texts make on their readers, through exposure to new vocabulary, different syntactic structures and deeper lexico-semantic networks, may in themselves encourage the development of new competencies and increase reader capacity to handle greater textual complexity.

We assume that it is the relationship with characters that a reader develops that is one of the important meaning-making processes in fiction reading. We examined the complex relationship between the character speech and eye-behaviour in children's fiction, its lexico-grammatical encoding and the links to children's "real-life" experience of similar situations. The language of 'eyes' is important in all three languages and in many respects, the three languages are very similar, for example, the main co-occurring verb types are similar, as are frequent implicit or explicit expressions of emotions, often encoded as a light metaphor in the eyes, with the eyes described as, for instance, shining, burning, flashing or twinkling.

In terms of characteristics and types of eye-behaviour and its narrative functions, we, again, find more similarities than differences between the languages compared – though ways of looking may be in subtle ways different, characters look into each other's eyes, or less frequently, avoid the gaze of the other. Eye-behaviour descriptions support the speech in highlighting the content or the manner of speaking in addition to, or instead of, reporting verbs, or they may even be used instead of a verbal response. Eye-behaviour descriptions can thus, to some degree, compensate for prosodic features and shared context that accompany non-fictional speech. Eye-behaviour in fictional worlds also performs specific additional functions in structuring the narrative and as a device contributes to the creation and development of the characters and the plot.

If fictional communication is to approximate communication in the "real" world, it cannot, therefore, be restricted to verbal code only. We hope to have shown that this applies to children's literature too, with eye-behaviour playing roles parallel to those children encounter in the non-fictional world – expressing emotions, communicating interpersonal attitudes, and accompanying and supporting speech. Just like in the "real" world, a gaze may be "polysemous", and its interpretation may depend not only on the relatively narrow context but also more generally on "real-life" knowledge and cultural background.

While the similar cultural background and the general narrative features lead to generally congruent types of eye-behaviour being employed in children's books in all three languages, the linguistic means available for the expression of eye-behaviour differ to a large extent, depending on the typological characteristics of the languages in question. The same semantic roles assigned to the eyes (such as the location or instrument) are often expressed by divergent syntactic means. The divergence reflects the relative weight of different word-order principles, the different means of indicating simultaneity, as well as the role of inflection in Finnish and Czech.

In terms of the syntactic encoding one of the differences that emerged from the analysis is greater dynamism, or agency, of eyes in English (the prevalence of the subject position): 'eyes' are much more frequently the 'doer' than in Czech and Finnish. For Finnish, the very precise encoding of 'location' stands out in comparison with English and Czech. This is perhaps not surprising considering the elaborate case system with a number of cases dedicated to expressing local relations. Another feature that emerged in terms of syntactic and grammatical encoding is greater use of non-verbal and participle (in Finnish E-inf.) constructions in Finnish and English that express the simultaneity of the process, that is, eye-behaviour descriptions are more frequently conceptualised as happening at the same time as the speech or other body language. In Czech non-finite verb forms are infrequent; instead, the temporal relations between speaking and looking are indicated by the verbal aspect, which marks the gaze either as an on-going process or as completed action or change of state of the eyes.

This was a pilot study on a limited sample. While the sample was sufficient to show the main tendencies in syntactic and functional distributions of fictional eye-behaviour descriptions across the three languages, larger data samples are needed for more fine-tuned lexical analysis as the lexical encoding seems to suggest subtle differences between the languages. The analysis of narrative strategies was likewise only limited. However, it showed potential for further interdisciplinary collaboration between linguists and literary scholars.

Acknowledgements

We would like to thank Dr. Hilkka Lindroos and two anonymous reviewers for their valuable comments and suggestions.

References

Alvstad, C. 2010. Children's Literature and Translation. In *Handbook of translation studies*, Vol. 1, Y. Gambier and L. van Doorslaer (eds), 22–27. Amsterdam: John Benjamins.

Argyle, M. 2010. *Bodily Communication.* Florence, UK: Taylor & Francis.

Beattie, G. 2004. *Visible Thought. The New Psychology of Body Language*. London: Routledge.

Brattico, P. 2020. Finnish Word Order: Does Comprehension Matter? *Nordic Journal of Linguistics*, 44(1), 38–70.

Brezina, V., Timperley, M. and McEnery, T. 2018. *LancsBox* (Version 4) [software]. Available from http://corpora.lancs.ac.uk/lancsbox [Last accessed 17 June 2021].

Burgoon, J.K., Guerro, L.K. and Floyd, K. 2010. *Nonverbal Communication*. London and New York: Routledge.

Čermáková, A. and Chlumská, L. 2017. Expressing 'place' in Children's Literature: Testing the Limits of the N-gram Method in Contrastive Linguistics. In *Cross-linguistic Correspondences. From Lexis to Genre*, T. Egan and H. Dirdal (eds), 75–96. Amsterdam: John Benjamins.

Čermáková, A. and Mahlberg, M. 2018. Translating Fictional Characters – *Alice* and *the Queen* from the Wonderland in English and Czech. In *The Corpus Linguistics Discourse*, A. Čermáková and M. Mahlberg (eds), 223–254. Amsterdam: John Benjamins.

Čermáková, A. and Mahlberg, M. Forthcoming. Gendered Body Language in Children's Literature Over Time. *Language and Literature.* Special issue edited by M. Burke and K. Coates.

Cvrček, V., Kodýtek, V., Kopřivová, M., Kováříková, D., Sgall, P., Šulc, M., Táborský, J., Volín, J. and Waclawičová, M. 2015. *Mluvnice současné češtiny 1. Jak se píše a jak mluví.* Univerzita Karlova v Praze: Karolinum.

Dušková, L. 2015. *From Syntax to Text. The Janus Face of Functional Sentence Perspective*. Univerzita Karlova v Praze: Karolinum.

Ebeling, S. O. 2014. An Eye for an Eye? Exploring the Cross-linguistic Phraseology of *Eye*/*Øye*. *Nordic Journal of Linguistics* 37(2), 225–255.

Firbas, J. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge: CUP.

Hakulinen, A. (ed.). 2004. *Iso suomen kielioppi*. SKS.

Hoffmannová, J. 1999. "Řeč očí" v konverzační analýze a interakční sociolingvistice. In *Dialog v češtině*, J. Hoffmannová and O. Müllerová (eds), 84–90. München: Verlag Otto Sagner.

Jerrim, J. and Moss, G. 2019. The Link between Fiction and Teenagers' Reading Skills: International Evidence from the OECD PISA Study. *British Educational Research Journal* 45(1), 181–200.

Johansson, M. and Håkansson, S. 2019. Corpus Stylistics and Literary Characterisation: Visual Moments in George Eliot. Abstract from 39th Annual Conference of the Poetics and Linguistics Association, Liverpool, United Kingdom.

Korte, B. 1997. *Body Language in Literature*. Toronto: University of Toronto Press.

Lambert, M. 1981. *Dickens and the Suspended Quotation*. New Haven, CT and London: Yale University Press.

Lathey, G. 2011. The Translation of Literature for Children. In *The Oxford Handbook of Translation Studies*, K. Malmkjær and K. Windle (eds), 198–213. Oxford: OUP.

Leech, G. and Short, M. 2007. [2nd ed.]. *Style in Fiction*. Harlow: Longman.

Lindquist, H. and Levin, M. 2008. Foot and Mouth: The Phrasal Patterns of Two Frequent Nouns. In *Phraseology: An Interdisciplinary Perspective*, S. Granger and F. Meunier (eds), 143–158. Amsterdam: John Benjamins.

Machálek, T. 2020. Kontext: Advanced and Flexible Corpus Query Interface. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 7003–7008.

Mahlberg, M., Wiegand, V. and Hennessey, A. 2020. Eye Language – Body Part Collocations and Textual Contexts in the Nineteenth-century Novel. In *Phraséologie et stylistique de la langue littéraire / Phraseology and Stylistics of Literary Language. Approches interdisciplinaires / Interdisciplinary Approaches*, L. Fesenmeier and I. Novakova (eds), 143–176. Berlin: Peter Lang.

Mahlberg, M., Wiegand, V., Stockwell, P. and Hennessey, A. 2019. Speech-bundles in the 19th-century English Novel. *Language and Literature*, *28*(4), 326–353.

Malá, M. and Šaldová, P. 2015. English Non-finite Participial Clauses as Seen Through their Czech Counterparts. *Nordic Journal of English Studies*, 14(1):232–257.

Mar, R.A. 2018. Stories and the Promotion of Social Cognition. *Current Directions in Psychological Science*, 27(4), 257–262.

McIntyre, D. 2016. Dialogue: Credibility Versus Realism in Fictional Speech. In *The Bloomsbury Companion to Stylistics*, V. Sotirova (ed.), 430–443. London: Bloomsbury.

Nádvorníková, O. 2020. Differences in the Lexical Variation of Reporting Verbs in French, English and Czech Fiction and their Impact on Translation. *Languages in Contrast* 20 (2), 209–234.

Nikanne, U. 2017. Finite Sentences in Finnish: Word Order, Morphology, and Information Structure. In *Order and Structure in Syntax I: Word Order and Syntactic Structure*, L.R. Bailey and M. Sheehan (eds), 69–97. Berlin: Language Science Press.

Nikolajeva, M. 2014. *Reading for Learning. Cognitive Approaches to Children's Literature.* Amsterdam: John Benjamins.

Oakhill, J., Cain, K. and Elbro, C. 2015. *Understanding and Teaching Reading Comprehension: A Handbook.* London: Routledge.

Page, N. 1988 [2nd ed.]. *Speech in the English Novel*. Houndmills: MacMillan.
Reynolds, K. 2011. *Children's Literature. A Very Short Introduction.* Oxford: OUP.
Šebestová, D. and Malá, M. 2019. Expressing Time in English and Czech Children's Literature: A Contrastive N-gram Based Study of Typologically Distant Languages. In *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium 2018*, J. Emonds, M. Janebová and L. Veselovská (eds), 469–483. Olomouc: Palacký University.
Sinclair, J. 2004. *Trust the Text. Language, Corpus and Discourse*. London and New York: Routledge.
Stephens, J. 2004. Linguistics and Stylistics. In *International Companion Encyclopaedia of Children's Literature*, P. Hunt (ed.), 99–111. London: Routledge.
Stockwell, P. and Mahlberg, M. 2015. Mind-modelling with Corpus Stylistics in David Copperfield. *Language and Literature*, *24*(2), 129–147.
Stubbs, M. 2007. Quantitative Data on Multi-word Sequences in English: The Case of the Word *World*. In *Text, Discourse and Corpora*, M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds), 163–189. London: Continuum.
Toivonen, I. 1995. *A Study of Finnish Infinitives* (Doctoral dissertation, Brandeis University).
Vaňková, I., Nebeská, I., Saicová-Římalová, L. and Šlédrová, J. 2005. *Co na srdci to na jazyku. Kapitoly z kognitivní lingvistiky*. Praha, Univerzita Karlova: Karolinum.
Wieçławska, E. 2012. *A Contrastive Semantic and Phraseological Analysis of the HEAD-related Lexical Items in Diachronic Perspective*. Rzeszow: Wydawnictwo Uniwersytetu Rzeszowskiego.
Wild, K., Kilgarriff, A. and Tugwell, D. 2013. The Oxford Children's Corpus: Using a Children's Corpus in Lexicography. *International Journal of Lexicography*, 26(2), 190–218.

Appendix

List of sources:

ENGLISH

(based on the BNC User Reference Guide, available at: http://www.natcorp.ox.ac.uk/docs/URG/bibliog.html)
Note: The BNC uses text extracts, the size of the extract is included, w.=words

A7A: 33,055 w. from *Bury the dead*. Carter, Peter. OUP, 1986
ABX: 36,224 w. from *Jubilee wood*. Hassall, Angela. OUP, 1989
AC4: 34,582 w. from *On the edge*. Cross, Gillian. OUP, 1989
AC5: 35,699 w. from *Paper faces*. Anderson, Rachel. Oxford University Press, 1991
ACB: 40,290 w. from *The lock*. Gates, Susan. OUP, 1990
ACV: 30,655 w. from *The forest of the night*. Kelly, Chris. OUP, 1991
AEB: 31,816 w. from *A twist of fate*. Scobie, Pamela. OUP, 1990
ALS: 4,149 w. from *Captain Pugwash and the huge reward*. Ryan, John. Gungarden Books Rye, 1991
AMB: 31,307 w. from *The adventures of Endill Swift*. McDonald, Stuart. Canongate Publishing Ltd, 1990
APW: 37,376 w. from *Quest for a babe*. Hendry, Frances Mary. Canongate Publishing Ltd, 1990
AT4: 44,592 w. from *Who, sir? Me, sir?* Peyton, K. M. OUP, 1988
B0B: 39,326 w. from *The Challenge book of brownie stories*. Moss, Robert. MTB Ltd, 1988
B2N: 907 w. from *How Miranda flew down Puddle Lane*. McCullagh, Sheila. Ladybird Books Ltd, 1991
BMS: 39,028 w. from *Gate-crashing the dream party*. Leonard, Alison. Walker Books Ltd, 1990
BMU: 38,121 w. from *The distance enchanted*. Gervaise, Mary. John Goodchild Publ., 1983
BPD: 27,335 w. from *Traffic*. Masters, Anthony Simon. Schuster Young Books, 1991
C85: 39,351 w. from *The first of midnight*. Darke, Marjorie. John Murray (Publishers) Ltd, 1989
CA3: 32,781 w. from *Lee's ghost*. Pulsford, Petronella. Constable & Company Ltd, 1990
CAB: 38,476 w. from *Goodnight Mister Tom*. Magorian, Michelle. Puffin Harmondsworth, 1983
CAX: 1,070 w. from *Polly and the privet bird*. Cartwright, Reg & Cartwright, Ann. Random House, 1992
CCA: 11,471 w. from *A bad spell for the worst witch*. Murphy, Jill. Puffin Harmondsworth, 1988
CE0: 3,400 w. from *Now then Davos*. Wiley, M., Harmer, D. & McMillan, I. Amazing Colossal Press, 1991
CFJ: 15,315 w. from *A tale of Anabelle Hedgehog*. Lawhead, Stephen. Lion Publishing, 1990
CJJ: 38,787 w. from *Space marine*. Watson, Ian. Boxtree, 1993
CM1: 36,658 w. from *High elves*. King, Bill & Chambers, Andy, Games Workshop, 1993
CM4: 39,412 w. from *Inquisitor*. Watson, Ian. Boxtree, 1993
EFJ: 40,024 w. from *Carrie's war*. Bawden, Nina. Puffin Harmondsworth, 1988

F99: 38,385 w. from *Adam's paradise*. Rush, A. Macmillan Publishers, 1989
FNS: 6,263 w. from *Alice in Wonderland: Oxford Bookworms ed*. OUP, 1993
FNY: 10,527 w. from *The Brontë story: Oxford Bookworms ed*. Vicary, Tim. OUP, 1991
FP5: 5,523 w. from *The coldest place on earth: Oxford Bookworms ed*. Vicary, Tim. OUP, 1992
FPE: 5,161 w. from *Dead Ma"s Island: Oxford Bookworms ed*. Escott, John. OUP, 1992
FPL: 6,203 w. from *The phantom of the opera: Oxford Bookworms edition*. Bassett, J. OUP, 1992
FPP: 6,645 w. from *Grace Darling: Oxford Bookworms ed*. Vicary, Tim. OUP, 1991
FPT: 5,760 w. from *Anne of Green Gables: Oxford Bookworms ed*. OUP.
FPU: 23,934 w. from *Great Expectations: Oxford Bookworms ed*. West, Claire. OUP, 1992
FPV: 15,297 w. from *Gulliver's travels: Oxford Bookworms ed.* OUP.
FR0: 37,862 w. from *The highest science*. Roberts, G. Virgin London, 1993
FR6: 31,194 w. from *Jane Eyre: Oxford Bookworms ed*. OUP, 1990
FRD: 6,558 w. from *Mary Queen of Scots: Oxford Bookworms ed.*. Vicary, Tim. OUP, 1992
FRE: 24,433 w. from *Far from the madding crowd: Oxford Bookworms ed*. West, C. OUP, 1992
FRK: 26,522 w. from *Oliver Twist: Oxford Bookworms ed*. Rogers, R. OUP, 1992
FRU: 10,658 w. from *Prisoner of Zenda: Oxford Bookworms ed*. Hope, A. & Mowat, D. OUP, 1993
FRX: 6,801 w. from *Robinson Crusoe: Oxford Bookworms ed*. Mowat, Diane. OUP, 1993
FS2: 10,645 w. from *The secret garden: Oxford Bookworms ed*. West, C. OUP, 1993
FS3: 9,090 w. from *The life and times of William Shakespeare: Oxford Bookworms*. Bassett, J. OUP, 1993
FSB: 8,817 w. from *The star zoo*. Gilbert, H. OUP, 1992
FSJ: 15,070 w. from *Treasure Island: Oxford Bookworms ed*. Escott, John. OUP, 1993
FSK: 8,178 w. from *Tooth and claw: Oxford Bookworms ed.* "Saki" Border, Rosemary. OUP, 1991
FSL: 4,773 w. from *Under the moon: Oxford Bookworms ed*. Akinyemi, Rowena. OUP, 1992
FSR: 39,905 w. from *White darkness*. McIntee, David. Virgin London, 1993
FUB: 16,637 w. from *The kingdom under the sea and other stories*. Aiken, Joan. Penguin Books, 1989
G1M: 38,774 w. from *Lucifer rising*. Mortimore, J Lane. A Doctor who books, 1993
GUS: 10,405 w. from *The picture of Dorian Gray: Oxford Bookworms ed*. Nevile, Jill. OUP, 1989
GV3: 6,140 w. from *The piano*. Border, Rosemary. OUP, 1989
GV7: 13,411 w. from *Dr Jekyll and Mr Hyde: Oxford Bookworms ed*. Border, Rosemary. OUP, 1991
GV9: 6,021 w. from *The love of a king*. Barnes, Trevor & Dainty, Peter. OUP, 1989
GVM: 6,047 w. from *New Yorkers*. Mowat, D. & Hutson, S. OUP, 1991 6-44
GW5: 8,765 w. from *Skyjack! Oxford Bookworms ed*. Vicary, Tim. OUP, 1989
GW8: 32,881 w. from *Tess of the d'Urbervilles: Oxford Bookworms ed.* West, Clare. OUP, 1989
GWA: 5,872 w. from *Voodoo Island*. Duckworth, Michael, OUP. 1989
GWC: 6,371 w. from *White death: Oxford Bookworms ed.* Vicary, Tim. OUP, 1989
GWH: 18,719 w. from *Wuthering Heights: Oxford Bookworms ed*. West, Clare. OUP, 1992
H0F: 39,022 w. from *The green behind the glass*. Geras, Adele. Lions Teen Tracks, 1989
H7V: 19,951 w. from *The hound of the Baskervilles: Oxford Bookworms ed*. Nobes, Patrick. OUP, 1989
H8G: 9,942 w. from *Frankenstein: Oxford Bookworms ed*. Nobes, Patrick. OUP, 1992
H8P: 6,365 w. from *Sherlock Holmes short stories: Oxford Bookworms edition*. West, Clare. OUP, 1989
H93: 1,586 w. from *The magician*. Escott, John. OUP, 1993
H9E: 2,154 w. from *Escape from Planet Zog*. Davies, Paul. OUP, 1992
H9U: 20,258 w. from *Ghost stories: Oxford Bookworms ed*. Border, Rosemary. OUP, 1989
HGS: 43,372 w. from *Frankenstein unbound*. Aldiss, Brian. New English Library Sevenoaks 1991
HTN: 35,729 w. from *A little lower than the angels*. McCaughrean, Geraldine. OUP, 1987
HTY: 41,032 w. from *The pit*. Penswick, Neil. Virgin London, 1993
CH0: 38,786 w. from *Krokodil tears*. Yeovil, Jack. GW Books Ltd, 1990
CH4: 39,631 w. from *Matilda*. Dahl, Roald. Puffin Harmondsworth, 1989
CH9: 6,850 w. from *The Minpins*. Dahl, Roald. Cape London, 1991
CHR: 10,339 w. from *Return of the red nose joke book*. Green, Rod. Boxtree, 1991

FINNISH

sla001: Annikki Marjala; Kaamosyön sankarit (1997)
sla002: Annikki Marjala; Korvatunturin salaisuus (1998)
sla003: Hannele Huovi; Salainen maa (1998)
sla004: Hannele Huovi; Tuliraja (1994)
sla005: Marja Luukkonen; Ihmeellinen omenatarha (1997)
sla006: Marja-Leena Tiainen; Jääprinsessa ja jäähykuningas (1996)
sla007: Tittamari Marttinen; Saaran taika (1996)

sla008: Laila Kohonen; Linnan uhka (1998)
sla009: Marja Luukkonen; Pikku Noita ja Karipeikko (1998)
sla010: Ritva Toivola; Turmankukka (1998)
sla011: Tittamari Marttinen; Seelan aurinkokello (1998)
sla012: Mari Lampinen (Kristina Carlson); Anni tien päällä (1998)
sla013: Mari Lampinen (Kristina Carlson); Annin uusi vuosi (1999)
sla014: Maria Vuorio; Matka, joka aina taittui (1996)
sla015: Mari Mörö; Sakun lintukesä (1998)
sla016: Leena Laulajainen; Sininen soittorasia (1998)
sla017: Tuija Lehtinen; Sara@crazymail.com (1998)
sla018: Taru ja Tarmo Väyrynen; Karri ja öiset valot (1998)
sla019: Laura Lähteenmäki; Rinkkadonna (1998)
sla020: Else Lassila; Korpin laulu (1999)
sla021: Anna-Liisa Haakana; Huityttö ja Pampoika (1999)
sla022: Sirpa Puskala; Pikkuruu Mustanmusta (1999)
sla023: Kari Levola; Sysimusta sukkapyykki (1999)
sla024: Heikki Willamo; Siiri Sopulin syksy (1998)

CZECH

Note: the year of publication indicates the edition included in the corpus, not necessarily the first year of publication

Batlička, Otakar (1979). *Tanec na stožáru*. Praha. Albatros.
Burdová, Michaela (2008). *Poselství jednorožců*. Praha. Fragment.
Čapek, Josef (2003). *Povídání o pejskovi a kočičce*. Praha. Albatros.
Čechura, Rudolf (2003). *Čítanka pro začínající detektivy*. Praha. Knižní klub.
Dědeček, Jiří (2013). *Jede jede klokan*. Praha. Dokořán.
Fischl, Viktor (1993). *Strýček Bosko*. Brno. Atlantis.
Flos, František (1987). *Lovci kožišin*. Praha. Volvox Globator.
Foglar, Jaroslav (1968). *Hoši od Bobří řeky*. Praha. Mladá fronta.
Franková, Hermína - Macourek, Miloš (1994). *Arabela*. Praha. Svoboda.
Háj, Felix (1990). *Školák Kája Mařík* [1.–3. díl] Praha. Vyšehrad.
Hlaváčková, Iva (2008). *Vládci Sedmihoří*. Praha. Fragment.
Hlaváčková, Iva (2008). *Terčina bláznivá dobrodružství*. Praha. Fragment.
Hlaváčková, Iva (2008). *Tajemná země Minor*. Praha. Fragment.
Hofman, Ota (1990). *Pan Tau a tisíc zázraků*. Praha. Albatros.
Hofman, Ota (1989). *Chobotnice z Čertovky*. Praha. Albatros.
Kabátová, Veronika (2006). *Když přijde láska*. Havlíčkův Brod. Fragment.
Kabátová, Veronika (2006). *Tenhle kluk je můj!* Havlíčkův Brod. Fragment.
Kabátová, Veronika (2006). *Nová láska na obzoru*. Havlíčkův Brod. Fragment.
Kárník, Zdeněk (2004). *Bílá ruka a poklad hradu Handštejna*. Praha. Dokořán.
Klimtová, Vítězslava (2006). *Obyčejný skřítek*. Kostelní Vydří. Karmelitánské nakladatelství.
Klimtová, Vítězslava (2006). *O statečném skřítku Drnovci*. Kostelní Vydří. Karmelitánské nakladatelství.
Klimtová, Vítězslava (2006). *Bukvínkova kouzelná píšťalka*. Kostelní Vydří. Karmelitánské nakladatelství.
Koutská, Blanka - Němcová, Božena (2000). *Sedmero krkavců a jiné pohádky*. Praha. Vyšehrad.
Kriseová, Eda (1992). *Terezka a Majda na horách* [1. část] Brno. Atlantis.
Kriseová, Eda (1992). *Terezka a Majda na horách* [2. část] Brno. Atlantis.
Kriseová, Eda (1992). *Terezka a Majda na horách* [4. část] Brno. Atlantis.
Kriseová, Eda (1992). *Terezka a Majda na horách* [3. část] Brno. Atlantis.
Kühnl, Daniel (2008). *Správná parta a zrušená kletba*. Praha. Fragment.
Kühnl, Daniel (2007). *Správná parta a kouzelný míč*. Praha. Fragment.
Macourek, Miloš (1971). *Pohádky*. Praha. Mladá fronta.
Macourek, Miloš (1982). *Mach a Šebestová*. Praha. Albatros.
Merhoutová, Eliška (2000). *Indické pohádky a bajky*. Praha. Vyšehrad.
Myslíková, Míla (2002). *V dobrém jsme se sešli...* Brno. L. Marek.
Neuwirth, Štěpán (2000). *Paseka živých jelenů*. Ostrava. Repronis.
Pavlíček, František (2003). *Princ Bajaja*. Brno. Atlantis.
Pavlíček, František (2003). *Tři oříšky pro Popelku*. Brno. Atlantis.
Pavlíček, František (2003). *Královský slib*. Brno. Atlantis.
Poláček, Karel (1967). *Bylo nás pět*. Praha. Odeon.

Pospíšilová, Zuzana (2006). *Moudrá sova Rozárka*. Havlíčkův Brod. Fragment.
Prášková, Markéta (2013). *Klobouky z Agarveny* 3. Praha. Grada.
Prášková, Markéta (2007). *Klobouky z Agarveny* 1. Havlíčkův Brod. Fragment.
Procházková, Helena (2009). *Mince krále Ašóky*. Praha. Baronet.
Renč, Ivan (2006). *Tajemství posledního večera*. Kostelní Vydří. Karmelitánské nakladatelství.
Rubík, Jan (2005). *O princezně Mirandolíně*. Kostelní Vydří. Karmelitánské nakladatelství.
Rudolf, Stanislav (1985). *Metráček*. Praha. Olympia.
Rudolf, Stanislav (1985). *Nebreč, Lucie*. Praha. Československý spisovatel.
Selucký, Oldřich (2004). *Pavel, dobrodruh víry*. Kostelní Vydří. Karmelitánské nakladatelství.
Semerád, Martin (2000). *Čarodějovy pohádky*. Praha. Autobus.
Skřivánek, Jaromír (1998). *Kouzelný hadí kámen*. Praha. Knižní klub.
Stanovský, Vladislav (2005). *Jak chodil Kristuspán se svatým Petrem po světě*. Kostelní Vydří. Karmelitánské nakladatelství.
Šmejkal, Roman (1996). *Zvířátková abeceda*. Praha. Knižní klub.
Troska, Jan Matzal (1997). *Peklo v ráji*. Praha. Leprez.
Vladislav, Jan (1999). *Pohádky paní Meluzíny*. Brno. Atlantis.
Vodňanský, Jan (1997). *Velký dračí propadák*. Praha. Volvox Globator.
Vopěnka, Martin (1998). *Pohádky Větrných hor*. Praha. Knižní klub.
Vyskočil, Ivan (1990). *Malý Alenáš*. Praha. Práce.
Wenig, Adolf (1997). *Pověsti o hradech*. Praha. Volvox Globator.
Werich, Jan (1977). *Fimfárum*. Praha. Albatros.
Unknown (1998). *Obrázky z Bible*. Praha. Vyšehrad.

Authors' addresses

Anna Čermáková
The Institute of the Czech National Corpus,
Faculty of Arts,
Charles University nám. Jana Palacha 1/2
CZ-116 38 Praha 1
Czech Republic
anna.cermakova@ff.cuni.cz

Markéta Malá
Department of English Language and Literature,
Faculty of Education,
Charles University,
Celetná 13,
CZ-116 39 Praha 1
Czech Republic
marketa.mala@pedf.cuni.cz

Pragmatic annotation of a domain-restricted English-Spanish comparable corpus

Rosa Rabadán[1], Noelia Ramón[1], Hugo Sanjurjo-González[2]

[1]University of León (Spain), [2]University of Deusto (Spain)

This paper explores the multi-layer annotation of a written domain-restricted English-Spanish comparable corpus (CLANES – Controlled LANguage English Spanish), focusing on pragmatic annotation. The annotation scheme draws on part of speech tagging and a semantic annotation scheme, i.e. the UCREL Semantic Analysis System, with some added categories to fit the food-and-drink domain represented in CLANES. These are used to build significant (pragmatic) metapatterns. Seven different pragmatic functions have been identified in our corpus, namely <STATE>, <DIRECT>, <SUGGEST>, <RECOMMEND>, <PRAISE>, <EVIDENCE> and <RELATE TO READER>. Computer scripts translate this linguistic information into regular expressions to be used in unsupervised annotation. Partial results indicate that applying lexical restrictors boosts the success rate considerably. However, metadata is preferred because of increased replicability and generality. Replicability issues and limitations encountered during testing are also addressed.

Keywords: semantic annotation, pragmatic annotation, comparable corpus, regular expressions, English/Spanish

1. Introduction

Richly annotated corpora are essential for the retrieval of usable information in different applied environments. In bilingual corpora, multi-layered annotation becomes vital to carry out detailed contrastive studies and as a basis for applications in the ever-increasing hybrid, human-machine text production flows. Most bilingual corpora feature at least part-of-speech (PoS) annotation, and some include other types of lexico-grammatical annotation, e.g. cohesive devices in Kunz and Lapshinova-Koltunski (2018), or genre-specific multiword combinations in Pizarro Sánchez (2017), among others. However, semantically and pragmatically annotated bilingual corpora are still rare and very much in demand.

This paper explores the multi-layer annotation of the domain-restricted English-Spanish comparable CLANES corpus (Controlled LANguage English Spanish), focusing primarily on pragmatic annotation. CLANES includes over 1.5 million words in the two working languages distributed across six subcorpora corresponding to two different written genres: informational-promotional texts of gourmet foods and drinks, on the one hand, and instructive texts in the

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.v11i1.3445.

form of culinary recipes, on the other. The aim is to automatically identify (semantic and) pragmatic meanings in such texts. As the initial tagging of the corpus was limited to PoS and rhetorical moves (Labrador *et al.*, 2014), an annotation scheme beyond the boundaries of the initial one needed to be developed. Thus, attempts at integrating semantic and pragmatic information had to be implemented to facilitate such analyses. Another limitation was the need to conduct unsupervised annotation at these levels on large amounts of texts.

In a traditional bottom-up approach to annotating textual material, we can find the first step in PoS tagging, namely assigning a particular grammatical category to each separate linguistic item. The next level in linguistic annotation, semantic annotation, will post a specific meaning to each item and multiword expression (MWE). And finally, pragmatic labels will refer to the speaker/writer's intentions when using the language for communicating with the intended audience, and contextual features play a role in the choice of linguistic elements. Considering the wide range of functions that may be performed employing language and all the extralinguistic factors involved, pragmatic annotation reveals itself as a complex task, more so if we try to carry it out automatically. As a result of these challenges, pragmatic annotation is still much less advanced than other linguistic annotation types.

Besides, most pragmatic studies are relatively small-scale qualitative analyses concentrating on spoken language data samples (Archer *et al.*, 2008: 613; Milá-García, 2018). Pragmatically annotated corpora of written texts are still rare but see Marín-Arrese's CESJD tagset (2017, 2019) or Weisser's in-progress TART dataset proposal (2018: 280ff).

A multi-layered linguistic analysis of CLANES may reveal a significant amount of relevant data for many applied purposes in the language industries, including teaching technical writing or developing semi-automatic applications for guided writing. With such applications in mind, the present paper describes the procedure followed for constructing a pragmatic annotation scheme to be applied to the CLANES corpus.

2. Data and method

2.1 Corpus description

This study's starting point is the CLANES corpus compiled at the University of León, Spain, in 2014-2019. It is a comparable corpus including 772,953 words in English and 776,100 in Spanish distributed across six subcorpora in each language. It comprises informational-promotional texts of gourmet foods and drinks and instructive texts in the same domain. The informational-promotional subcorpora include texts on wine (López Arroyo and Roberts, 2016), cheese (Labrador and Ramón, 2020), biscuits, herbal teas (Izquierdo and Pérez-Blanco, 2020) and dried meats (Ortego Antón, 2020). The instructive subcorpus is made up of culinary recipes (Rabadán *et al.*, 2016). All the texts were retrieved from online company web pages, open-access blogs and producer/retailer-facilitated materials. Table 1 shows the number of words per language in each subcorpus.

Table 1. Number of words in the CLANES corpus.

CLANES CORPUS		
Name of subcorpus	**Number of words English**	**Number of words Spanish**
RECIPES	290,498	257,184
CHEESE	128,347	139,017
WINE	117,874	140,694
BISCUITS	98,994	81,456
DRIED MEATS	85,419	42,161
HERBAL TEAS	51,821	115,588
TOTAL	**772,953**	**776,100**

2.2 Method and working procedure

Initially, the CLANES corpus was PoS tagged using TreeTagger (Schmid, 1995) and rhetorically annotated using an *ad hoc* tool, the ACTRES Tagger.[1] The long-term aim underlying the annotation project was to develop support for authors of promotional texts in the food and drinks industry (Labrador and Ramón, 2020). It soon became evident that the use of the annotated materials initially was limited to contrastive rhetoric and grammatical analyses and that attempts to go beyond these boundaries required higher-level semantic and pragmatic information (see this section and section 3 below).

The semantic annotation scheme employed to tag all the words in this corpus was based on USAS (UCREL Semantic Analysis System, Rayson *et al.*, 2004) developed at the Lancaster University from 2013, covering several different languages, including English and Spanish. The USAS scheme is based on the Longman Lexicon of Contemporary English (McArthur, 1986), composed of 21 major discourse fields and 232 labels, shown in Table 2 (based on Archer *et al.*, 2002).

Table 2. Twenty-one major discourse fields of the USAS scheme.

A – General and abstract terms	**B** – The body and the individual	**C** – Art and crafts
E - Emotion	**F** – Food and farming	**G** – Government and public
H – Architecture, housing and the home	**I** – Money and commerce in industry	**K** – Entertainment, sports and games
L – Life and living things	**M** – Movement, location, travel and transport	**N** – Numbers and measurement
O – Substances, materials, objects and equipment	**P** – Education	**Q** – Language and communication
S – Social actions, states and processes	**T** – Time	**W** – Word and environment
X - Psychological actions, states and processes	**Y** – Science and technology	**Z** – Names and grammar

Due to the general nature of the USAS categories, the semantic annotation had to be implemented manually by adding more specific subcategories from the F domain, which is highly relevant to the CLANES material, e.g. F1: Food has been expanded into F1.1, accounting for 'variety/class of food x,' such as *jamón ibérico* (< Iberico ham). F1.2 marks 'meal organization', i.e., when the food is typically eaten, as this is an important cross-cultural difference: *breakfast, dinner, snack*. F1.3 indicates 'cuts,' i.e., meat/ fish commercial cuts such as *chop, steak, fillet*, etc. The resulting semantic dataset includes over 5,000 domain-specific entries in both languages, plus an additional 10,000 general language entries in Spanish.

The amount of data spiked the need to conduct unsupervised annotation at these levels on larger amounts of text. Manual tagging was effected on a section of the corpus using first regular expressions and symbolic analysis. Then, a semantic word labelling tool (Sanjurjo-González, 2020) that includes different NLP (Natural Language Processing) processes together with word2vec (Mikolov *et al.*, 2013) and fastText (Bojanowski *et al.*, 2017) algorithms were used for unsupervised annotation. Current semantic annotation results show an overall degree of success of 89% in Spanish, including MWEs. For English, the success rate is close to 90%. These results refer to the food-and-drinks domain dataset in CLANES.

Pragmatic annotation starts from identifying combined PoS and semantic patterns that indicate one particular pragmatic function (see section 3 below). Our initial scheme includes

[1] Rhetorical move tagger® Available at http://contraste2.unileon.es/web/en/tagger.html. ACTRES stands for Contrastive Analysis and Translation English-Spanish in its Spanish acronym (Análisis Contrastivo y Traducción English-Spanish) https://actres.unileon.es/.

six categories, namely <STATE>, <DIRECT>, <SUGGEST>, <RECOMMEND>, <PRAISE> and <EVIDENCE>. An additional category, <RELATE TO READER>, was identified when testing replicability on popular science texts (see section 4 below).

<STATE> simply marks the delivery of referential information and applies to names of products, dishes, etc., as in *Buxton Blue*, *Hafner Vineyards 2009 Chardonnay*. <DIRECT> indicates an action to be carried out to fulfil a goal, as in *stir into batter* or *remove from oven*. <RECOMMEND> singles out the best course of action for the task at hand, as in *best eaten at room temperature*. <SUGGEST> signals that options offered may or may not be put into practice, as in *it can be enjoyed all year round* or *food pairing suggestions*. <PRAISE> refers to the product's good properties, as perceived intersubjectively, as in *perfectly balanced flavour combination*. <EVIDENCE> adds positive factual information about the product, as in *this cheese has won many awards.* <RELATE TO READER> promotes and marks the reader's involvement in the text, as in *I will save you, reader, the detailed account of …; but what does 'thermal equilibrium' really mean? I refer the reader to 1.15, where …*

Once the tagset had been defined, one of the first issues we had to address was segmentation. How could we decide where to set the boundaries of our pragmatic annotation scheme? Most previous attempts at (automatic) pragmatic annotation are applied to spoken data. However, the CLANES corpus contains written texts, although with very specific contextual settings: promotional and instructive texts from the food and drink industry. Bearing in mind that we were dealing with written texts, segmentation based on turns was not an option. It was decided to employ full stops and other punctuation marks indicating sentence boundaries as the 'pragmatic unit' to be tagged, as "all 'semantically complete' units, even if they consist of syntactic fragments (e.g. single noun phrases (NPs) that answer questions), should have a meaning and pragmatic function that is largely independent of the surrounding meanings and is thus also worth labelling individually" (Weisser, 2015: 89).

2.3 Developing the CLANES Annotation Scheme

The CLANES pragmatic annotation scheme uses the Python programming language and consists of two independent scripts that perform two primary tasks. The first script is responsible for converting the patterns into valid regular expressions using re-Python package syntax. Patterns are rule-like and are used to locate a particular combination of meta-items within sentence-based strings. The second script's role is to match those regular expressions with tokens of a specific, pragmatic function. Roughly, it works as follows: Texts are segmented into sentences using the NLTK sentence tokenizer (Bird *et al*., 2009). The script runs the regular expressions through the segmented units and checks whether and where a match can be found. If so, it applies the corresponding pragmatic tag (Figure 1).

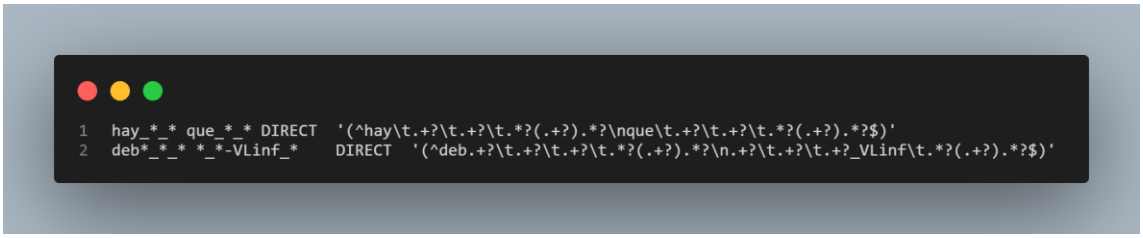

Figure 1. Python regular expressions sample.

Pattern querying is done through the ACTRES Corpus Manager (Sanjurjo-González, 2017). This custom-made user platform allows for retrieving information for three layers of metadata, at present grammatical and semantic, and pragmatic (this last one still under construction).

Figure 2 shows the browser interface, where the searches can be carried out by lemmas or by PoS and semantic categories. More than one semantic category can be listed. In Figure 2, we have selected a search for any verb in the base form with the semantic category F1 food, followed by a determiner and followed by any noun in the singular with the semantic category F1 food.

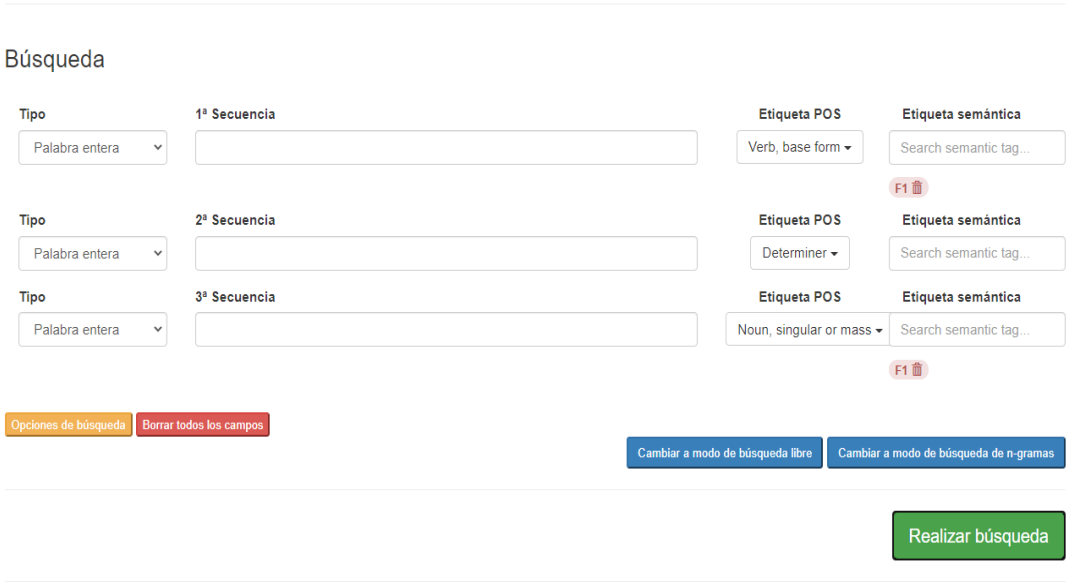

Figure 2. Interface of the ACTRES Corpus Manager.

Figure 3 shows some of the hits for the query described in Figure 2 above: *roll the potato mixture, combine the flour, roll each pancake,* etc. We can see that all the concordance lines are displayed with all their PoS and semantic tags for further analysis.

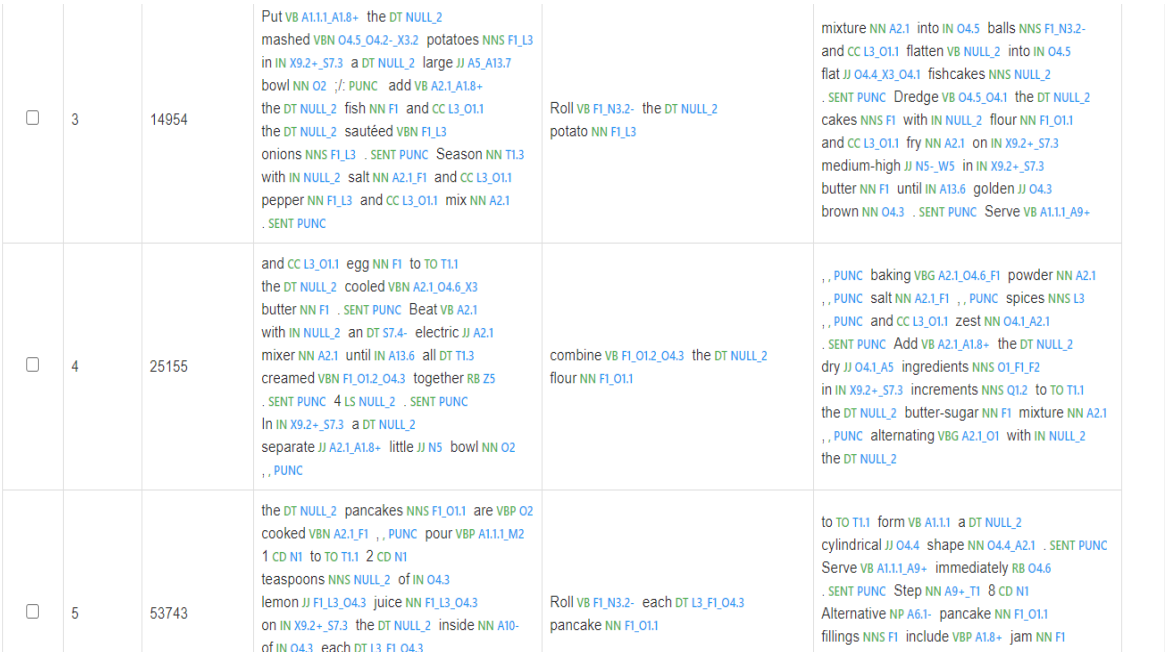

Figure 3. Sample hits of the ACTRES Corpus Manager.

3. Results: CLANES pragmatic annotation

This section describes the pragmatic patterns designed using combined strings of the PoS and semantic tags in the CLANES corpus of food and drinks, as illustrated in Figures 2 and 3. Preliminary results of unsupervised annotation testing are also included (section 3.6). The aim was to obtain prototypical patterns that could be used to identify each of the pragmatic functions described above. Computer scripts would be successfully applied to annotate the corpus with pragmatic tags.

3.1 Pragmatic function <DIRECT>

The pragmatic function <DIRECT> refers to an action to be carried out to fulfil a goal. To build the patterns that will lead to *regular expressions*, we identified both obligatory and optional pattern components combining PoS and semantic tags. Starting manually from PoS tags, it was found that the use of a verb in the imperative form (VB) was the most common mark of <DIRECT> in English. It was commonly followed by some noun (NN, NNS) in the field of food or drink (*add onions*), a field-related object (*remove the pan from the stove*) or a time expression with cardinal numbers (*simmer 6 to 7 minutes*). Optionally, determiners, prepositions, adverbs, adjectives or past participles may appear in between the obligatory items of verb and noun, as in *stir in salt and pepper; serve slightly chilled; enjoy with milk; serve on cheeseboard, etc.*

We listed all the semantic labels attached to the obligatory building blocks in this pattern to boost pattern identification further (see Table 2 for an overview of the semantic/discourse fields represented by A, F, O, etc.). Verbs: A1.1.1, general actions; A1.8, inclusion/exclusion; A2.1, affect: modify, change; A9, getting and giving: possession; A10, open/ closed, hiding/hidden, finding, showing; F1, food; F2, drinks; M2, putting, taking, etc.; O4.6, temperature, and X3, sensory; and nouns occurring after the verb: F1, food; F2, drinks; F4, farming and horticulture; L3, plants; O2, objects generally. The optional components have been shown to work better if only their PoS tags were considered determiners (DT), prepositions (IN), adjectives (JJ) or adverbs (RB), as well as cardinal numbers (CD) or past participles (VBN).

An example of a text chunk correctly tagged as <DIRECT> reads as follows:

(1) <DIRECT> *Stir* VB A1.1.1_M1 *in* INX 9.2 *salt* NNA2.1_F1 *and* CC *pepper* NNF1_L3 </DIRECT>.

Table 3 shows the <DIRECT> pattern in English with the obligatory components highlighted.

Table 3. <DIRECT> pattern in English.

PATTERN <DIRECT>					
OBLIGATORY		[OPTIONAL 1]	[OPTIONAL 2]	OBLIGATORY	
PoS	USAS			PoS	USAS
VB	A1.1.1	DT	CD	*NN*	F1
	A1.8	IN	JJ		F2
	A2.1	RB	VBN		F4
	A9				L3
	A10				T1.3
	F1				O2
	F2				

	M2				
	O4.6				
	X3				
	X3.3				

In English, additional restrictions need to be implemented in the search to distinguish imperatives from the infinitives, as they are not inflected, and both tagged as VB, verb base form. These restrictions include leaving out instances of VB preceded by an item labelled semantically as *Z5* (*grammatical bin*), or with PoS tags *pronoun PP*, or *noun NN*, as in (2) where 'you PP I2.2' signals that 'remove' is not an imperative.

(2) When WRB O4.4 you PP I2.2 remove VB A2.1_A1.8 the DT Z5 beans NNS F1_L3

In Spanish, however, these restrictions were not required, as the PoS annotation marks verbal inflections. Alternatives in the verbal slot are a *se*-passive (*se añaden la sal y la pimiento* (add (the) salt and pepper)) or a first-person plural present (*añadimos la sal y la pimienta* (we add (the) salt and pepper).

The same procedure was replicated for all other pragmatic patterns in English and Spanish to construct a pragmatic annotation scheme. The patterns have been translated into regular expression rules and subsequently used to carry out an unsupervised pragmatic annotation.

3.2 Pragmatic function <RECOMMEND>

The pragmatic function <RECOMMEND> indicates the best course of action to savour, prepare or present a food or drink item among those available and represented across all CLANES subcorpora. We have identified two typical patterns. Tables 4 and 5 show the <RECOMMEND> pattern in English with the obligatory and the optional components. The first (a) includes an adjective semantically tagged as A5.1+ Evaluation: good/bad; O4.2 Judgement of appearance; X3.1+ Sensory: taste, or X3.5 Sensory: smell followed by a preposition or, less frequently, by a condition marker, as in *excellent with grilled red meats*.

Table 4. <RECOMMEND a> pattern in English.

<RECOMMEND a> PATTERN			
OBLIGATORY		OBLIGATORY	
PoS	USAS	PoS	USAS
JJ	A5.1+	*IN*	Z5
	O4.2	IF	Z7
	X3.1+		
	X3.5		

The second pattern (b) features a superlative (RBS) followed by a preposition (IN) or an adverb (RB) is also possible in this function, as in *best at six months of ageing*. The adverb must semantically belong to X9.2 Ability: Success and failure or S7.3 Competition as *in best served slightly cool at 12-15ºC*. Optionally, this pattern may include a past participle (VBN) belonging to one of the following semantic categories: A1.1.1 General actions, making; A4.1 Generally kinds, groups, examples; E2 Liking; S3.1 Relationship: General, as in *best enjoyed with milk*.

Table 5. <RECOMMEND b> pattern in English.

<RECOMMEND b> PATTERN					
OBLIGATORY		[OPTIONAL]		OBLIGATORY	
PoS	USAS	PoS	USAS	PoS	USAS
RBS	A5.1	VBN	A1.1.1	*IN*	Z5
	A13.2		A4.1	*RB*	X9.2
	X3.1		E2		S7.3
	X3.5		S3.1		
			Z7		

Examples of text chunks tagged as <RECOMMEND> read as follows:

(3) <RECOMMEND> excellent A5.1_O4.2+ with Z5 grilled F1 red O4.3 meats F1 </RECOMMEND>

(4) <RECOMMEND> best RBS A13.2_A5.1+++ served VBN A1.1.1_A9 at IN Z5 room NN A1.1.1_A10 temperature NN O4.6 </RECOMMEND>.

In the case of Spanish, we find the same <RECOMMEND a> pattern: *ideal para ensaladas* (ideal in salads); *perfecto para tus picoteos* (perfect as a snack). Additionally, in Spanish, we also have a reflexive passive (A6.2, Q2, F1 and F2) followed by an infinitive or an adjective without a specific semantic profile, as in *se recomienda acompañar de un vino blanco Generoso* (we recommend pairing it with a Generous white wine).

3.3 Pragmatic function <SUGGEST>

The pragmatic function <SUGGEST> offers alternatives to carry out the task that may or may not be put into practice, and it appears across all subcorpora. We identified a primary pattern (Table 6) consisting of a pronoun (PP) or a noun (NN), whose meaning falls in the domain food (F1), drinks (F2) or L3 (plants), followed by a modal (MD) indicating possibility, a verbal base form and a past participle. The latter needs to be semantically tagged as A1.1.1 General actions, making; A2.1 Affect: Modify, change or E2 Liking, as in *it can be served hot* or *onions may be cooked in advance*.

Table 6. <SUGGEST a> pattern in English.

OBLIGATORY		OBLIGATORY		OBLIGATORY		OBLIGATORY	
PoS	USAS	PoS	USAS	PoS	USAS	PoS	USAS
PP	Z8	*MD*	A7	*VB*	A3	*VBN*	A1.1.1
NN	F1						A2.1
	F2						E2
	L3						

We were also able to single out a secondary pattern (Table 7) using an *–ing* form (VBG) of verbs meaning A1.1.1 general actions; A9 getting and giving: possession, or F1 food, combined with a noun (NNS) meaning Q2.1 Speech: Communicative or Q2.2 Speech acts, as in *serving suggestions* or *(food) pairing suggestions*.

Table 7. <SUGGEST b> pattern in English.

<SUGGEST b> PATTERN			
OBLIGATORY		OBLIGATORY	
PoS	USAS	PoS	USAS
VBG	A1.1.1	*NNS*	Q2.2
	A9-		Q1.1
	F1		

Examples of text chunks tagged as <SUGGEST> read as follows:

(5) <SUGGEST>mango NN L3_F1 can MD A7 be VB A3 replaced VBN A2.1 with IN Z5 sugar NN F1</SUGGEST>

(6) <SUGGEST> Food NN F1 pairing NN F1 suggestions NNS Q2.2 </SUGGEST>.

In Spanish, this pragmatic function's main pattern involves a modal verbal periphrasis with *se: se puede sustituir por leche de almendras* (it can be replaced with almond milk). An alternative is using the 1st person plural in the modal verb followed by an infinitive: *podemos utilizar jengibre en polvo* (we may use ginger powder). Additionally, we found a pattern similar to English <SUGGEST b>: a noun (Q2.2) in the plural optionally followed by a preposition: *Sugerencias de degustación* (serving suggestions); *maridaje* (food pairing).

3.4 Pragmatic function <PRAISE>

The pragmatic function of <PRAISE> refers to the product's good properties, as perceived intersubjectively. This function is widespread in the promotional texts of our corpus. In this case, we identified a pattern with three elements with the following PoS tags: one obligatory (a positive adjective JJ) and two optional elements: a pre-modifying adverb (RB) and a noun (NN-NNS) placed after the adjective. Both optional items may occur at the same time: *wonderfully creamy texture; truly lovely cheese*. At least one of the optional elements must co-occur with the adjective: either a pre-modifying adverb (*intensely fruity; absolutely delicious*) or the noun being pre-modified (*delicious milk; toasty aroma*).

The obligatory adjective in this pattern must belong to one of the following semantic categories: A.12 Easy/difficult; A5.1 Evaluation: good/bad; O4.2 Judgement of appearance; O4.3 Colour and colour patterns; O4.5 Texture; T2 Time: beginning and ending; T3 Time: old, new and young; age; X3.1 Sensory: taste; X3.3 Sensory: touch; X3.5 Sensory: smell. Moreover, the noun being modified must belong to one of the following semantic categories: A1.8 Inclusion/exclusion; A5.1 Evaluation: good/bad; F1 food; F2 drinks; X3.1 Sensory: taste; X3.5 Sensory: smell. The pre-modifying adverb may belong to any semantic category.

An example of a text chunk correctly tagged as <PRAISE> reads as follows:

(7) <PRAISE>wonderfully RB creamy JJ X3.1_O1.1 texture NN O4.5 </PRAISE>.

Table 8 shows the <PRAISE> pattern in English with the obligatory components highlighted.

Table 8. <PRAISE> pattern in English.

PATTERN <PRAISE>				
[OPTIONAL 1]	**OBLIGATORY**		[OPTIONAL 2]	USAS
	PoS	**USAS**		
RB	**JJ**	A12	NN	A1.8
		A5.1		A5.1
		O4.2		F1
		O4.3		F2
		O4.5		X3.1
		T2		X3.5
		T3		
		X3.1		
		X3.3		
		X3.5		

In Spanish, the pragmatic function of <PRAISE> is very similar and also pivots around an obligatory adjective with significantly positive semantic tags (A5.1: Evaluation: good/bad; X3.1: Sensory: taste; O4.2: Judgement of appearance), preceding or following a common noun (NC in the Spanish PoS notation system) with one of the following semantic tags: A5.1, F1, F2, X3.1 and X3.5. An example tagged for <PRAISE> reads as follows:

(8) <PRAISE>Es VSfin A3+_L1_Z5_X2.4 un ART Z5_N1_T3_T1.2_Z8 bizcocho NC F1 muy ADV A13.3 esponjoso ADJ O4.5_O4.1</PRAISE> (it is a fluffy, delicious sponge cake).

3.5 Pragmatic function <EVIDENCE>

The pragmatic function of <EVIDENCE> adds positive factual information about the product, e. g. comments about medals or awards won by the product, Protected Designation of Origin or other quality certifications in the food and drinks domain: *a gold medal winner at the World Cheese Awards.* Together with <PRAISE>, the pragmatic function of <EVIDENCE> is widespread in promotional discourse. The intended audience will be more inclined to buy a particular product if it has certified proof of quality.

In this case, we have noticed that several different patterns pivot around one single semantic label, namely S7.3: Competition, whether this semantic label is attached to an adjective (JJ): *an award-winning cheese*; to a noun (NN/NNS) such as *medal* or *winner*, or to a verb (VB), such as *award* or *win*, as in *this celebrated cheese has won many medals*; *this tea has won many awards*. The various PoS strings in which these items engage are typical unmarked syntactic patterns of English, such as adjective + noun or verb + determiner + noun. These common patterns are abundant and singled out as evidence by the sole presence of the semantic tag S7.3.

An example of the function <EVIDENCE> reads as follows:

(9) <EVIDENCE> A DT Z5 gold JJ O4.3 medal NN O2_S7.3 winner NN X9.2_S7.3 at IN Z5 the DT Z5 World NN W1 Cheese NN F1 Awards NN S7.3 </EVIDENCE>.

Similar patterns with the dominance of the S7.3 semantic category were observed in Spanish:

(10) <EVIDENCE> Ganador NC X9.2_S7.3 del PDEL Z5 premio NC S7.3 Cincho NP O2_A6.2 de PREP Z5 Oro NC O1 2006</EVIDENCE>. (Winner of the Cincho de Oro 2006 award).

We also tested these patterns on texts from different domains. In the case of popular science materials, we found the same pattern. However, the noun's semantics refers to someone in an authorial position: S7.2 Respect, P1 Education, X9.1 Ability, intelligence, as in *Professor at Princeton University; winner of the Nobel Prize.*

3.6 Other pragmatic functions: <RELATE TO READER> and <STATE>

The pragmatic function of <RELATE TO READER> signals the author's willingness to seek and maintain the reader's involvement. It is a phatic function, also found in academic lectures (Hyland 2005: 182–189), aiming to ensure the reader's engagement in text flow and progression (see section 2.2). Relating to the reader is done by addressing the reader directly, using rhetorical questions to mark advancement in presenting facts or concepts. The typical pattern consists of an interrogative such as *how* or a *wh*-pronoun/ adverb (WRB/WP) followed by an interrogative clause, which is uninformative in the context.

Another strategy, which can be combined with the one just mentioned, addresses the reader directly. An example of the function <RELATE TO READER> reads as follows:

(11) <RELATE TO READER> But CCB Z5 how WRB Z5 does DZ Z5_ A1.1.1 dark JJ W2 energy NN1 Y1_W1_X5.2+ work VB I3.1_A1.1.1? SENTPUNC </RELATE TO READER>.

A parallel pattern has been noted in Spanish, consisting of an interrogative clause featuring the standard word order required by Spanish syntax, i.e., ¿*Qué tienen que ver, podría el lector preguntar, estas cuestiones de biología y de química con la uniformidad del universo primitivo*? (What, the reader might ask, do these questions of biology and chemistry have to do with the early universe's uniformity?).

The pragmatic function of <STATE> can be considered a default category that describes products or narrates sequences of actions and may adopt an almost unlimited combination of elements. This situation results in great difficulty in operationalizing patterns for this particular function. Any pragmatic segment not assigned to any of the other functions will be considered <STATE> as in *All the milk is unpasteurized.*

3.7 Preliminary testing results

The ACTRES pragmatic annotation scheme has been repeatedly tested at different stages of its development. Preliminary results showed a score of around 75% in Spanish and 62.5% in English. If taken by subcorpus, the informational-promotional subcorpora's success rate exceeded 70% in Spanish and was near 60% in English. For the instructive genre (recipes), the overall results hit 84% in Spanish and just below 65% in English. If taken by pragmatic category, in Spanish, the success rate ranged from 92% for <RECOMMEND> to 43.44% for <SUGGEST>. In English, the accuracy ranged from 88% for <STATE> to 5% for <SUGGEST>. These trials were all carried out using lexical (content word) restrictors in addition to PoS and semantic categories. For example, a sentence including the Spanish verbal periphrasis *hay que* (have to) would be identified automatically as having the pragmatic function <DIRECT>, or the noun *award* would prompt the tag <EVIDENCE>. Annotation testing exclusively using metadata (with no lexical restrictors) is currently underway.

4. Limitations and replicability

Results so far demonstrate that pragmatic tagging of written texts faces several challenges. Deciding on the unit for pragmatic annotation and the relevant segmentation of the text was one of them. Using punctuation marks such as full stops as sentence boundaries seemed to be the right choice initially. However, texts in our corpus contain promotional language, including titles, headings, and subheadings, most of which are not followed by any punctuation mark that could be used to define unit boundaries. As a result, some of our segments include more than one illocution, as headings tend to be grouped with the first sentence after the header. This means that the script only assigns one pragmatic function where manual annotation would assign two. Apart from manual revision, the possible solution is to consider other typographical features to discriminate segments adequately, e.g., paragraph indents.

Other limitations spring from mistagging in the PoS or semantic layers, which is misleading when identifying the patterns that form the basis for the regular expressions script. Minor but time-intensive manual corrections have been necessary at both levels to ensure that any minor mistake or null tag in a particular linguistic unit will not interfere with accurate pragmatic tagging. In our case, this problem has been an issue, as the PoS tagset makes use of different tags for English and Spanish, our two working languages. Both the searches and the pattern formulation are carried out using language-specific tags, which considerably slows down the process. Upgraded versions of the browser will try to achieve homogeneity in this respect.

Our tagset suffers from underspecification, particularly in the pragmatic function <STATE>, requiring a more detailed design to stop being the "default" function.

Replicability is central in annotation schemes. We have run informal pattern recognition tests in popular science (Rabadán and Gutiérrez-Lanza, 2020) and business texts (Pizarro 2017; Rabadán *et al.* in press). The goal was to check whether the patterns triggering the regular expressions hold in different domains and genres. Business texts revealed that <EVIDENCE> and <STATE> are typically found in reports; <RECOMMEND> was also found although marginally. Our test on popular science materials yielded a massive output of <STATE>, occasional but regular cases of <EVIDENCE>, and an additional pragmatic function that had not materialized in CLANES, but was added as a result of this test, <RELATE TO READER>.

5. Conclusions and further work

The long-term aim of setting up a pragmatic annotation scheme is to offer essential support to authors/communicators in the food and drinks industry. Previous attempts in this and other environments (Labrador and Ramón, 2020; Rabadán *et al.*, in press) showed the need for better, more informative corpora. Annotation has been a staple feature of written corpora for decades now but is still mainly confined to part-of-speech tagging. Although indisputably useful, additional information types become essential when facing tasks more sophisticated than basic grammatical contrast. Semantic annotation mostly follows USAS with some domain-specific additions. We aimed to set up a pragmatic annotation scheme based on previous PoS and semantic information. Using both layers, PoS and semantic, we identified prototypical patterns that have been used to characterize seven pragmatic functions. A computer script transformed the patterns into regular expressions, whose role is to detect tokens of a particular pattern within a sentence. Another script executes the unsupervised annotation using the regular expressions.

Our tests have shown that using lexical restrictors in the patterns boosts the success rate considerably. However, it detracts significantly from cross-linguistic replicability since sets of lexical restrictors would have to be changed according to language, genre, and domain. For

example, verb forms tagged grammatically as infinitives will tend to be functioning pragmatically as <directive>, and semantic categories like A5.1+ (Evaluation: good) will always be tagged as <PRAISE> in any text type/domain, not only in promotional discourse in the food and drink industry. With nouns, however, semantic categories may need a different dataset according to particular domains, i.e., F1: Food nouns (e.g. salad, *ensalada*) would be replaced by Y1: Science and technology in a popular science corpus (e.g. spiral galaxy, *galaxia espiral*) or I2: Business in a business reports corpus (e.g. assets, *activos*).

The tests also suggest that pragmatic function frequency is linked to text type and overall text function rather than the domain. Results highlight the need for robust metapatterns rather than lexical items as pattern restrictors. They further suggest that adding an additional layer of annotation with more detailed information on grammatical functions would improve the usefulness of "supporting metadata" for pragmatic annotation. An example is verbal periphrases, which contribute meanings unrepresented in current PoS tags, for example, aspect types, such as inchoative in *poner a hervir* (start boiling) or continuative and gradual in *ir añadiendo* (roughly, keep adding) (Yllera, 1999: 3412–3420).

Replicability depends on the metapatterns underlying the regular expressions that allow the computer scripts to "extract rules" and apply them successfully to corpus annotation. So far, our pragmatic categories seem to work outside their home domain of food and drink.

Work in progress focuses on streamlining the regular expressions to improve script performance and success rate in pragmatic tagging and upgrading browser capabilities. Results will enable a wealth of studies and contribute to developing new applications, such as building a pre-editing workbench for bilingual text production of instructive and promotional genres in the food-and-drink domain. They will also improve existing author support tools (Rabadán *et al.*, in press) and be an essential component in designing a "drafter controlled language" for specific domains.[2]

Acknowledgements

This research has been funded by grant FFI2016- 75672-R awarded by the Spanish Ministry of Science and Innovation and ERDF (European Regional Development Fund).

References

Archer, D., Wilson, A and Rayson, P. 2002. Introduction to the USAS Category System. Retrieved from: http://ucrel.lancs.ac.uk/usas/usas_guide.pdf [Last accessed 21 March 2021].

Archer, D., Culpeper, J. and Davies, M. 2008. Pragmatic Annotation. In *Corpus Linguistics: An International Handbook*, M. Kytö and A. Lüdeling (eds), 613–642. Berlin: Mouton de Gruyter.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media http://www.nltk.org/book_1ed/ [Last accessed 26 November 2020].

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5: 135–146.

Hyland, K. 2005. Stance and Engagement: A Model for Interaction in Academic Discourse. *Discourse Studies* 6 (2): 173–191. DOI: 10.1177/1461445605050365.

[2] A "drafter controlled language" is a restricted interactive language that presents stub sentences and guides the user to gradually complete these sentences as required using the lexicogrammatical, semantic and pragmatic information on offer (Kuhn 2014: 138).

Izquierdo, M. and Pérez Blanco, M. 2020. A Multi-level Contrastive Analysis of Promotional Strategies in Specialised Discourse. *English for Specific Purposes* 58: 43-57. DOI: 10.1016/j.esp.2019.12.002.

Kuhn, T. 2014. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics* 40(1): 121–170. DOI: 10.1162/COLI_a_00168.

Kunz, K. and Lapshinova-Koltunski, E. 2018. English vs. German from a Textual Perspective: Looking inside Chain Intersection. *Bergen Language and Linguistics Studies* 9(1): 21–42. DOI: 10.15845/bells.v9i1.1520.

Labrador, B., Ramón, N., Alaiz-Moretón, H. and Sanjurjo-González, H. 2014. Rhetorical Structure and Persuasive Language in the Subgenre of Online Advertisements. *English for Specific Purposes* 34: 38–47. DOI: 10.1016/j.esp.2013.10.002.

Labrador, B. and Ramón, N. 2020. Building a Second-language Writing Aid for Specific Purposes: Promotional Cheese Descriptions. *English for Specific Purposes* 60: 40–52. DOI: 10.1016/j.esp.2020.03.003 9.

López Arroyo, B. and Roberts, R.P. 2016. Differences in Wine Tasting Notes in English and Spanish. *Babel* 62 (3): 370–401. DOI: 10.1075/babel.62.3.02lop.

Marín Arrese, J. 2017. Multifunctionality of Evidential Expressions in Discourse Domains and Genres. Evidence from Cross-linguistic Case Studies. In *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-Pragmatic Perspectives*, J. Marín Arrese, G. Hassler and M. Carretero (eds), 195–224. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/pbns.271.

Marín Arrese, J. 2019. CESJD-JMA Tagset for Annotation of Epistemic and Effective Stance Markers [Data set]. http://corpusnet.unileon.es/assets/uploads/tools/CESJD-TAGSET.pdf [Last accessed 1 June 2021].

McArthur, T. 1986. *Longman Lexicon of Contemporary English*. London: Longman.

Mikolov, T., Chen, K., Corrado, G., and Dean, J.. 2013. Efficient Estimation of Word Representations in Vector Space. In arXiv preprint arXiv:1301.3781v3. Ithaca: Cornell University.

Milá-García, A. 2018. Pragmatic Annotation for a Multi-Layered Analysis of Speech Acts: A Methodological Proposal. *Corpus Pragmatics* 2: 265–287.

Ortego Antón, M.T. 2020. Las fichas descriptivas de embutidos en español y en inglés: un análisis contrastivo de la estructura retórica basado en corpus. *Revista Signos* 53 (102): 170–194.

Pizarro Sánchez, I. 2017. A Corpus-based Analysis of Genre-specific Multiword Combinations. Minutes in English and Spanish. In *Cross-linguistic Correspondences: From Lexis to Genre*, T. Egan and H. Dirdal (eds), 221–252. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/slcs.191.09san.

Rabadán, R., Colwell, V. and Sanjurjo-González, H. 2016. BiTeXting Your Food: Helping the Gastro Industry Reach the Global Market. In CILC2016 (EPiC Series in *Language and Linguistics*, vol. 1), A. Moreno Ortiz and C. Pérez-Hernández (eds), 361–371. https://easychair.org/publications/open/Wv4r; https://doi.org/10.29007/4xtp [Last accessed 1 June 2021].

Rabadán, R. and Gutiérrez-Lanza, C.. 2020. Developing Awareness of Interference Errors in Translation: An English-Spanish Pilot Study in Popular Science and Audiovisual Transcripts. In *Specialised Languages and Multimedia. Linguistic and Cross-Cultural Issues*, E. Manca and F. Bianchi. Special issue of *Lingue e Linguaggi*, 40: 379–404. http://siba-ese.unisalento.it/index.php/linguelinguaggi [Last accessed 1 June 2021].

Rabadán, R., Pizarro, I. and Sanjurjo-González, H. In press. Authoring Support for Spanish language Writers: A Genre-restricted Case Study. *Revista Española de Lingüística Aplicada*, RESLA.

Rayson, P., Archer, D., Piao, S. and McEnery, T. 2004. The UCREL Semantic Analysis System. In *Proceedings of the Workshop Beyond Named Entity Recognition, Semantic Labelling NLP Tasks* (LREC 2004), 7–12. Lisbon: European Language Resources Association.

Sanjurjo-González, H. 2017. *Creación de un Framework para el tratamiento de corpus lingüísticos* [Development of a Framework for corpus linguistic analysis]. Doctoral dissertation, University of León, Spain.

Sanjurjo-González, H. 2020. Increasing Accuracy of a Semantic Word Labelling Tool Based on a Small Lexicon. In *Proceedings of the 17th International Conference on Natural Language Processing* (ICON-2020), Patna, India.

Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.
Weisser, M. 2015. Speech Act Annotation. In *Corpus Pragmatics: A Handbook*, K. Aijmer and C. Rühlemann (eds), 84–113. Cambridge: Cambridge University Press.
Weisser, Martin. 2018. *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/scl.84.
Yllera, A. 1999. Las perífrasis verbales de gerundio y participio. In *Gramática descriptiva de la lengua española*, I. Bosque and V. Demonte (eds), 3391–3441. Madrid: Espasa.

Authors' addresses

Rosa Rabadán
Department of Modern Languages
Campus de Vegazana
University of León
ES-24071 León
Spain
rraba@unileon.es

Noelia Ramón
Department of Modern Languages
Campus de Vegazana
University of León
ES-24071 León
Spain
noelia.ramon@unileon.es

Hugo Sanjurjo-González
Department of Computing, Electronics and Communication Technologies
University of Deusto
Faculty of Engineering
Unibertsitate Etorbidea 24006
ES-48014 Bilbao
Spain
hugo.sanjurjo@deusto.es