UCCTS-4

Using Corpora in Contrastive and Translation Studies

Lancaster University

24th to 26th July 2014

Abstract Book

Plenary presentations

Intermodal corpora in contrastive and translation studies

Silvia Bernardini
University of Bologna
silvia@sslmit.unibo.it

Intermodal corpora, i.e., corpora featuring parallel or comparable bilingual texts produced under different translation modalities (typically oral and written), constitute a relatively new but very promising source of data, not just for translation, but for contrastive studies as well. In this talk I first survey the sources of data commonly used in corpus-based contrastive and translation studies, and argue that intermodal corpora afford a novel perspective that can enrich both fields. I then briefly describe EPTIC, the European Parliament Translation and Interpreting Corpus, which builds on the well-known EPIC (European Parliament Interpreting Corpus), and makes available independently produced translational and interpretational outputs based on input from the European Parliament plenary sessions, as well as the input source texts/discourses themselves. The corpus, still under construction, is thus intermodal and bidirectional (English <=> Italian).

To illustrate the potential of this type of corpus, I present data on the use of collocations in interpreted and translated English. The comparison highlights quantitative and qualitative similarities and differences whose implications are discussed both in a contrastive and translational perspective. I conclude by arguing that EPTIC could and should be enlarged through an exercise of collaborative corpus construction, and make some suggestions as to how we, as a research community, could proceed.

.

Contrastive phraseology: method and analysis

Signe Oksefjell Ebeling
University of Oslo
s.o.ebeling@ilos.uio.no

In this talk I will outline a method applied in the analysis of patterns in contrast (Ebeling & Ebeling 2013), where patterns are defined as recurrent word-combinations with semantic unity. The contrastive approach is inspired by scholars who advocate translations and cross-linguistic correspondences as *tertium comparationis* (e.g. James 1980, Altenberg 1999, Johansson 1998, 2007) and Chesterman's (1998, 2007) concept of perceived similarity. Bidirectional translation data play an important role in this respect. The focus on patterns (phraseology) is inspired by the observation that meaning, to a greater extent than is often believed, is said to reside in multi-word units rather than single words. These units, or patterns, and Sinclair's (1996, 1998) extended-unit-of-meaning model are therefore central to the approach. To illustrate the method, a short case study will be presented. The study shows that patterns weave an intricate web of meanings across languages and demonstrates the need for more phraseology-oriented contrastive studies.

References

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

Chesterman, Andrew. 1998. *Contrastive Functional Analysis*. Amsterdam: John Benjamins.

Chesterman, Andrew. 2007. Similarity analysis and the translation profile. In W. Vandeweghe, S. Vandepitte & M. van de Velde (eds). *The Study of Language and Translation*. *Belgian Journal of Linguistics* 21. 53–66.

Ebeling, Jarle & Signe O. Ebeling. 2013. *Patterns in Contrast*. Amsterdam: John Benjamins.

James, Carl. 1980. *Contrastive Analysis*. Longman.

Johansson, Stig. 1998. On the role of corpora in cross-linguistic research. In S. Johansson & S. Oksefjell (eds). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 3–24.

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

Sinclair, John. 1996. The search for units of meaning. *Textus IX. 75–106.*

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

Food for thought: metaphor, machines and translation

Dorothy Kenny
Dublin City University
dorothy.kenny@dcu.ie

This paper explores how the rise of contemporary machine translation stands to change how we understand translation itself. Drawing on corpus-based approaches to metaphor and metonymy, as well as existing studies of metaphors of translation, it focuses in particular on how computer scientists construct translation in a one-million word corpus of scholarly articles published in the journal *Machine Translation* between 2003 and 2013 and asks what such constructions might mean for translation and translators.

Contrastive linguistics as a discovery procedure

Béatrice Lamiroy
KU Leuven
beatrice.lamiroy@arts.kuleuven.be

Although at its origin (the sixties of last century, e.g. Alatis 1968, Weinreich 1953), contrastive linguistics was primarily associated with applied linguistics, aiming in particular at making foreign language teaching and learning more efficient, it has been defined in more recent years as a sub-discipline of linguistics with theoretical implications (e.g. König & Gast 2007, Lauridsen & Lauridsen 1989, Lahousse et al. 2010). This is the perspective that will be adopted in this talk.

The central hypothesis I will advocate for is that comparative linguistics is a discovery procedure, i.e. comparing languages contributes to a better understanding of linguistic forms and functions, not only on the usual assumptions of linguistic typology, but also for the individual languages. Thus, like any comparative practice, contrastive linguistics heeds light on similarities and differences, but it has a particular heuristic value in that it yields findings which are difficult to reach by the separate study of single languages.

Although both typologists and contrastive linguists basically assume that languages do not vary randomly nor without limits, they differ in methodology. Whereas typologists usually compare a large sample of languages with respect to a single property, e.g. modality (Van der Auwera & Plungian 1998), constrastive linguistics rather compare two languages, but extensively (e.g. König & Gast 2007, Van Belle & al. 2010).

After a general introduction on the topic of contrastive linguistics, the above mentioned hypothesis will be illustrated by a test case, viz. a series of French connectives (*en fait, de fait, en effet* and *en réalité*) which will be analysed in contrast to their Dutch equivalents on the basis of a parallel corpus French-Dutch.

References

Alatis, J. (ed.) 1968. *Contrastive Linguistics and its Pedagogical Implications*. Washington: Georgetown University Press.

König, E. & Gast, V. 2007. *Understanding English-German contrasts*. Berlin: E. Schmidt Verlag.

Van Belle, W., Lamiroy, B. Van Langendonck, W., Lahousse, K., Lauwers, P. & Van Goethem, K. 2010. *Een Nederlandse Grammatica voor Franstaligen*. http://wwwling.arts.kuleuven.be/NGF_N/NGF_NL.htm

Van der Auwera, J. and Plungian, V. 1998. Modality's Semantic Map. *Linguistic Typology* 2. 79-124.

Weinrieich, U. 1953. *Languages in Contact*. *Findings and Problems*. New York: Linguistic Circle of New York.

Beyond translation properties: The contribution of corpus studies to empirical translation theory

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

The study of translation properties is probably the best studied area in corpus-based translation studies. And while it is firmly situated within descriptive translation studies, it has not yet led to the development of an empirically-informed translation theory as proposed by Toury (1995, 2004).

This paper reviews the achievements of the corpus-based approach to translation and discuss some current related research questions. Many studies in corpus-based translation research revolve around describing the specific properties attributed to translation as summarised in Baker (1993), often concentrating on discussing corpus frequencies of individual features. As claimed by Tummers et al. (2005) in the context of cognitive linguistics, this approach to corpus research limits the range of potential more general explanations or even predictions. Feature frequencies certainly play an important role for establishing the empirical facts of translation, but considering the complex interplay of factors in the context of translation, they may not be sufficient to develop a theory of translation, the eventual goal of empirical translation studies. In my view, two areas are particularly promising in terms of explaining and predicting the outcome of the translation process.

First, the use of multivariate statistics should allow us to take the corpus approach one step further. This means specifically not just testing the statistical significance of individual feature frequencies in comparison to some reference but rather accounting for the (cumulative) effect various features have simultaneously. I will illustrate this area with ongoing work on visualising hidden patterns in the CroCo Corpus of aligned English – German source and target texts (Hansen-Schirra et al. 2012) based on relative frequencies of 28 lexico-grammatical features (Evert and Neumann 2013).

Secondly, the need to bring closer together the two strands of empirical translation research, namely process-based and corpus-based research, has already been pointed out in particular by Halverson (2013), Alves et al. (2010). One of the main sources of explanation for translation properties is the translator's understanding of the source text during translating. However, isolating this as a cause for characteristic properties of translation products in corpora is problematic. Product-based studies

therefore need to be complemented by studies of the translation process geared specifically to testing hypotheses about causes of translation properties. I will exemplify the link between process and corpus data with ongoing research on applying corpus methods to recorded translation process data.

While an attempt at empirically modelling translation is not yet within reach, I propose that studies in these areas further our understanding of the inner workings of translation which will ultimately enable translation scholars to develop a theory of translation based on empirical evidence.

References

Alves, F., A. Pagano, S. Neumann, E. Steiner, and S. Hansen-Schirra. 2010. "Units of Translation and Grammatical Shifts: Towards an Integration of Product- and Process-Based Research in Translation." In Translation and Cognition, edited by G. Shreve and E. Angelone, 109–142. Amsterdam: Benjamins.

Baker, M. 1993. "Corpus Linguistics and Translation Studies. Implications and Applications." In Text and Technology. In Honour of John Sinclair, edited by M. Baker, G. Francis, and E. Tognini-Bonelli, 233–250. Amsterdam: Benjamins.

Evert, Stefan, and Stella Neumann. 2013. "The Impact of Translation Direction on the Characteristics of Translated Texts: A Multivariate Analysis for English and German." Workshop "New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies" at the 46th Annual Meeting of the Societas Linguistica Europaea, Split, 18-21 September 2013.

Halverson, S. L. 2013. "Implications of Cognitive Linguistics for Translation Studies." In Cognitive Linguistics and Translation Advances in Some Theoretical Models and Applications. Berlin, Boston: De Gruyter.

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

Tummers, J., K. Heylen, and D. Geeraerts. 2005. "Usage-based Approaches in Cognitive Linguistics: A Technical State of the Art." Corpus Linguistics and Linguistic Theory 1 (2): 225–261.

Toury, G. 1995. Descriptive Translation Studies and Beyond. Amsterdam: Benjamins.

Toury, G. 2004. "Probabilistic Explanations in Translation Studies. Welcome as They Are, Would They Qualify as Universals?" In Translation Universals. Do They Exist?, edited by A. Mauranen and P. Kujamäki, 15–32. Amsterdam: Benjamins.

Paper presentations

The design and use of a translation corpus

Sumie Akutsu
J. F. Oberlin University
smakutsu@obirin.ac.jp

Tim Marchand
J. F. Oberlin University
marchand@obirin.ac.jp

1 Introduction

This paper discusses the design and rationale of creating a bilingual translation corpus from the writings of university students in Japan. The paper explains how the corpus, consisting of source texts in the learners' L1 (Japanese) and students' efforts at translating the texts into English, can be utilised to provide lesson materials for future groups of students and to find common errors among the Japanese learners.

These common errors can often be attributed to learners' tendency to translate sentences word by word without due consideration to the relevant meaning in context (Akutsu, 2009; 2010), and the paper explores how lesson materials with translation activities targeting certain common errors can be used to raise awareness of the pitfalls of direct translation. Elicitation data is used to determine the efficacy of this approach, and the paper concludes by arguing that a similar translation corpus may be of benefit to groups of learners from different L1 backgrounds.

2 Background

One of the difficulties in the field of English language teaching and learning in Japan is to communicate in writing (McKinley, 2006; 2010). In the case of Japanese university students, one of the major barriers to effective communication through writing is the tendency to translate directly from their L1. As Japanese learners are usually trained to do sentence by sentence translation based on some particular grammatical points or functions, they tend to think about what to say in Japanese first then try to translate it word by word directly from Japanese into English without interpreting the relevant meanings in context and without realizing the awkwardness in the resulting expressions (Cook, 2012).

The average Japanese student in university is typically much more coherent and expressive in Japanese than English; therefore, it is natural for them to struggle to put their advanced Japanese into simple English. In order to raise awareness of the fact that a language has a culture behind it, and word-by-word translation between two languages or cultures is not always possible, it is important to encourage students to avoid direct translation. Even though translation has been criticized under the trend of communicative approach (Cook, 2010), the Common European Framework of Reference for Languages defines translation both as an effective means of language learning and as a mediation skill in today's globalized world.

3 Rationale for translation activities

Through creative translation of Japanese prose into English, students have been shown to improve their writing ability in English while raising their language and cultural awareness (Snell-Hornby, 1995).

Based on the fact that the majority of Japanese university students confess that they think in Japanese and then try to translate into English, a model of a three-stage system of translation by Eugen Nida is proposed to facilitate creative writing in a feedback session (Munday, 2001). This process involves analyzing the structure of the source language, transferring it into the translation process, and restructuring it into a natural expression. The aim of this is to reproduce the intention of the original text, rather than trying to reproduce literally accurate text. The desired outcome is that students become more conscious of learning strategies in the study of English, and thus become dexterous in the use of these translation strategies.

According to Friedlander (1990), the positive effect of first language and translation usage in writing is "not just to generate content but also generate and verify appropriate word choice" (p.111). This view is supported by Laufer, who has demonstrated the pedagogical advantage using translation activities to improve learners' awareness of natural collocations (Laufer & Girsai, 2008). Using a strategy of first language reference is therefore expected to enhance learner writing in English. Direct translation is a habit the student needs to break, but guidance on the correct usage of dictionaries and references can help form new, constructive habits. Even when language exposure is limited and no instructor is around, students should be able to guide themselves to the best possible conclusion. Through this exercise, their awareness of cultures and language will be raised, and this can contribute to their further development as an effective language learner and thus a user.

4 The Translation Corpus

While we would argue that the judicious use of translation activities in the language classroom are of pedagogic value in themselves, a further benefit

can be derived from collating the learner texts to form the translation corpus. The translation corpus can be used in two ways. In order to cultivate their writing style, learners can compare and analyse texts of translated works by native and non-native speakers of English from the corpus. Through realizing cultural differences between Japanese and English languages in this way, students will possibly be more prepared to become autonomous language learners with better communication strategies.

Secondly, the learners' contribution to the translation corpus can be analysed like any other learner corpus, with common errors pinpointed and reified. The paper will demonstrate some examples of these common errors, and how their elucidation then informed the design of subsequent translation activities.

References

Akutsu, S. (2009). Creative writing: Using a translation exercise to improve students' writing skills – part I. *The Journal of Rikkyo University Language Center*, 21, 3- 10.

Akutsu, S. (2010). Creative writing: Using a translation exercise to improve students' writing skills – part II. *The Journal of Rikkyo University Language Center*, 24, 3-11.

Cook, G. (2010). *Translation in Language Teaching*. Oxford: Oxford University Press.

Cook, M. (2012). Revisiting Japanese English teachers' (JTEs) perceptions of communicative, audio-lingual, and grammar translation (yakudoku) activities: Beliefs, practices, and rationales. *The Asian EFL Journal*, 14(2), 79-98.

Grenfell, M. & Harris, V. (1999). *Modern Languages and Learning Strategies: In Theory and Practice*. London: Routledge.

Laufer, B. & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, 29(4), 694-716.

McKinley, J. (2006). Learning English writing in a Japanese university: Developing critical argument and establishing writer identity. *The Journal of Asia TEFL*, 3(2), 1-35.

McKinley, J. (2010). English language writing centres in Japanese universities: What do students really need? *Studies in Self-Access Learning Journal*, 1(1), 17-31.

Munday, J. (2001). *Introducing Translation Studies: Theories and Applications*. London: Routledge.

Snell-Hornby, M. (1995). *Translation Studies: An Integrated Approach*. Amsterdam: J. Benjamins.

Teaching, learning and translating Italian collocations through learner corpus

Marilei Amadeu Sabino
UNESP – São José do Rio Preto – Brazil
amadeusm@ibilce.unesp.br

Learner corpus research (LCR) stands at a crossroads among some disciplines as corpus linguistics, second language acquisition, foreign language teaching, and the results of the investigations conducted in this area may bring benefits to several research fields, namely, lexicography, contrastive linguistics, teaching methodology, cognitive linguistics, second language acquisition, foreign language teaching, language testing, natural language processing and translation.

Collocations are one of the several types of phraseologisms and although a lot has already been done in terms of phraseological research, it still remains a lot to be done in terms of extracting, describing, defining, teaching and learning these structures.

Granger et al. (2002, p. 7) argue that computer learner corpora are "[…] electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose". A very significant advantage of learner corpora is the fact that the researcher can have a record of the learners' production which may enable him to report what learners actually produce in terms of phraseological patterns.

Altenberg and Eeg-Olofsson (1990), Sinclair (1991), Fontenelle (1994), Granger (1998), Orenha-Ottaiano (2004; 2012), Meunier and Granger (2008) claim that the learning of collocations and other prefabricated chunks is crucial to learners who aim to produce fluent speech and they assert that the use of corpora in the foreign language classrooms promotes the teaching of these chunks. Thus, based on the well-known importance of providing students with the ability to use these prefabricated structures well, we built a parallel learner corpus made up of students' translations from Portuguese into Italian language. Therefore, this paper aims at showing some results of an investigation carried out in a Brazilian public university with students that attend a translation course.

The subjects of this research are university students from the 3rd year of a B. A. in Translation Course, whose level of Italian varies from intermediate to upper-intermediate. The original texts that comprise the corpus are newspaper articles taken from very popular Brazilian newspapers and magazines. The typology of the texts is related to

current world news and the topics selected were "One year after Tsunami in Japan"; "Financial crises in Greece and in Europe"; "Unemployment"; "Elections in the US"; "Bullying"; "Abortion", etc. These texts originally written in Portuguese were translated into Italian by a group of 10 students. With the help of *WordSmith Tools* (Scott 2004), it was possible to extract the data and analyse students' collocations.

The methodology of this investigation, corpus design and compilation are based on a similar research carried out by Orenha-Ottaiano (2012) in the same university, with the same translation students, the same original Portuguese texts, but translated into English.

Our aim is to compare, in a second stage, the collocations used by the Brazilian learners of Italian to the ones employed by the Brazilian learners of English, in order to check if:

a) Brazilian learners of English and Italian as foreign languages have the same difficulties in producing collocations;

b) they produce similar collocational errors; and

c) there is some kind of influence of the mother tongue on their choices.

Some of the problems found in the translation from Portuguese to Italian are related to the following collocations: "cessar fogo", "travar combates", "máxima autoridade rebelde", "governo transitório", "medidas de prevenção", "chegar ao poder", "zona do euro", "cobrir os empréstimos", "pacote de cortes", "rombo fiscal", to name a few.

For example, as learners are usually influenced by their mother tongue (Portuguese), they translated the collocation "entrevista coletiva" into "conferenza collettiva", when they should have used "conferenza stampa". And by ignoring the frequently used collocation "derrubou a resistência" in Italian, they translated it into "ha rovesciato la resistenza", "ha annullato la resistenza", "ha fatto cadere la resistenza", instead of into "ha piegato la resistenza".

The investigation allowed us to observe the students' collocational choices and patterns; the influence of the mother tongue on these choices; the most frequent collocational errors produced; and the most/least used type of collocations employed by them.

As a result of their production, we recognize the importance of teaching and encouraging students to explore the potential benefits of using corpora in translation. We also argue that when the teaching of collocations is in a more explicit (or intentional) way, it brings more benefits to learners than in the cases teachers hope it happens automatically, i. e., in an implicit (or incidental) way. As previously mentioned, the results of this research will be compared to Orenha-Ottaiano's findings and further discussed in a paper.

References

Altenberg, B.; Eeg-Olofsson, M. 1990. "Phraseology in Spoken English: presentation of a Project". In: AARTS, J.; MEIJS, W. (Ed). Theory and practice in Corpus Linguistics. Amsterdam: Randpi, p. 1-26.

Fontenelle, T. 1994. "Towards the construction of a collocational database for translation students". Meta 39 (1), p. 47-56.

Granger, S. 1998. *Learner English on computer*. London/ New York: Longman.

Granger S.; Hung, J.; Petch-Tyson, S. (Ed.) 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam/Philadelphia: John Benjamins.

Meunier, F.; Granger, S. 2008. "Phraseology in foreign language learning and teaching. Where to and from?" In: MEUNIER, F.; GRANGER, S. (Ed.). Phraseology in foreign language learning and teaching. Amsterdam: John Benjamins, p. 247-252.

Orenha-Ottaiano, A. 2004. *A compilação de um glossário bilíngüe de colocações, na área de jornalismo de negócios, baseado em corpus comparável*. Master's thesis, Universidade de São Paulo, São Paulo.

Orenha-Ottaiano, A. 2012. "English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy". *Acta Scientiarum*, 34 (1), p. 241-251.

Sinclair, J. 1991. *Corpus, concordance and collocation*. Oxford: Oxford University Press.

Thomas, J. E. (forthcoming). "Stealing a march on collocation". *TALC 10 Procceedings*.

WordSkew: Tracking the distribution of words and phrases within texts

Michael Barlow
University of Auckland
mi.barlow@auckland.ac.nz

1 Introduction

Corpus studies have benefitted from and relied upon software tools such as concordancers. The KWIC format has proved to be a simople but powerful form of display of textual data that enables subtle patterns to be revealed. However, the highlighting of some components of text data, such as collocational sequences, inevitably backgrounds other aspects of the texts and discourse.

The usefulness of KWIC format necessarily entails a focus on local patterns, which are exhibited without explicit reference to sentence or discourse boundaries, for example. Thus concordance-based analyses are the most part text-structure neutral.

2 WordSkew

WordSkew takes a different tack and starts with text structure and then moves on to look at lexicogrammatical patterns associated with text structure: sentences, paragraphs, or other units defined by the user. We know that words or phrases are not uniformly distributed within a text. What we don't know is how the clustering of words relates to text structure.

The *skew* in *Wordskew* refers to the assumption that the more interesting patterns of distribution of words or phrases across sentences or paragraphs or other text units will not be uniform but biased towards beginnings or middles or ends of the text unit. Thus the core function is to obtain a frequency profile of a word or phrase across different units in a text: the sentence, paragraph, section, and text as a whole.

3 A basic example

Figure 1 illustrates this skewing with the not very surprising example of the distribution of *however* within sentences taken from a corpus of British newspaper articles.

The data shows the marked preference for *however* in sentence-initial position, defined here as the first 10% of the sentence. The information is presented in two forms: a histogram and the table. Further information is given at the top of the screen and we can see that there are about 28 million words in the corpus and around 16,000 instances of *however*.

It is also possible to plot the distribution of words by position in the sentence: first word, second word, etc.

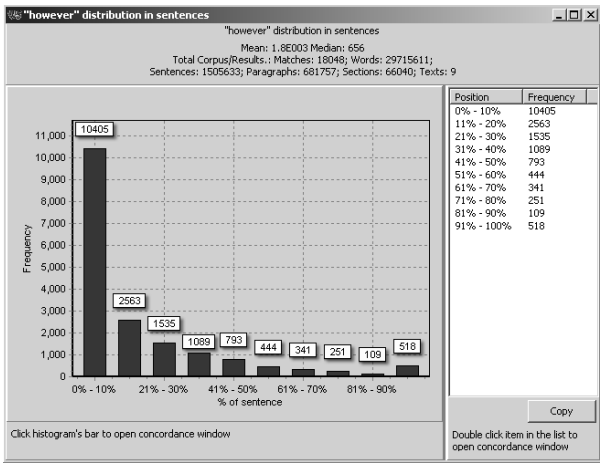

Figure 1: Distribution of *however* in sentences

In this example a simple search for the word *however* was initiated, with the empirical data adding details to our intuitions about a bias towards sentence-initial uses. We can contrast this sentence pattern with the distribution of *however* in paragraphs and in newspaper articles (here rendered as sections).

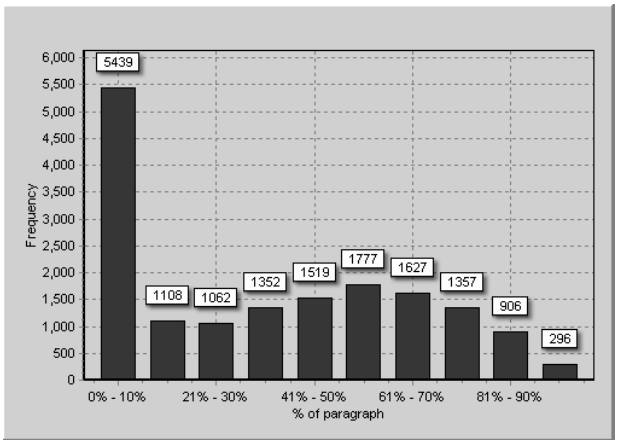

Figure 2: Distribution of *however* in paragraphs

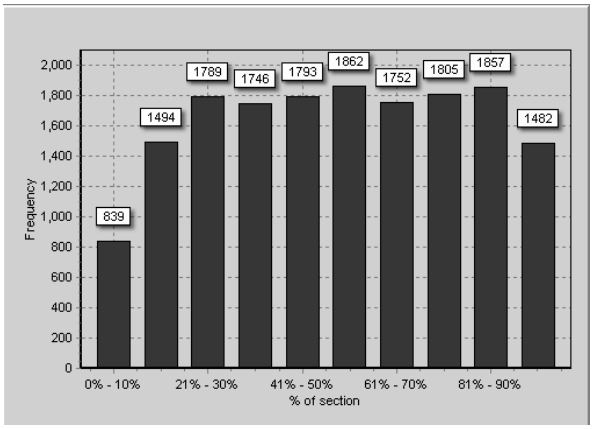

Figure 3: Distribution of *however* in articles

The article has been divided into ten units and, perhaps not surprisingly we find that *however* is less likely to occur in the initial part of the article compare with the remainder.

4 Another simple example

Figure 4 shows the distribution of the phrase a move in sentences. Once we obtain the data in relation to position within a text unit such as the sentence, it is possible to get the concordance lines for a particular position, as shown in Figure 5.

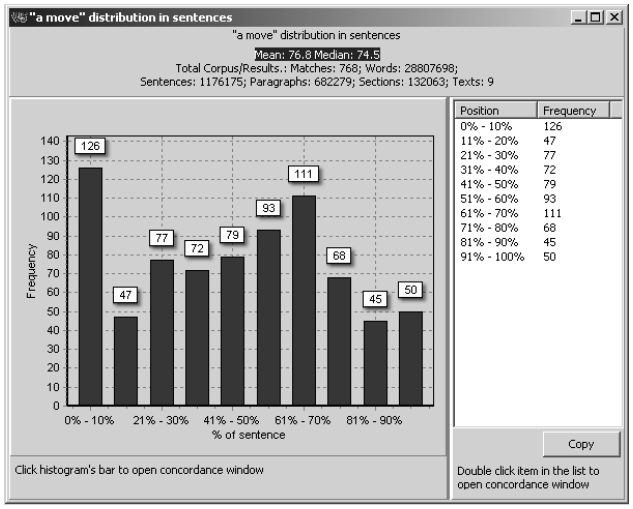

Figure 4: The distribution of *a move* in sentences

Hence the relation to text segments is primary and then concordance data is examined.

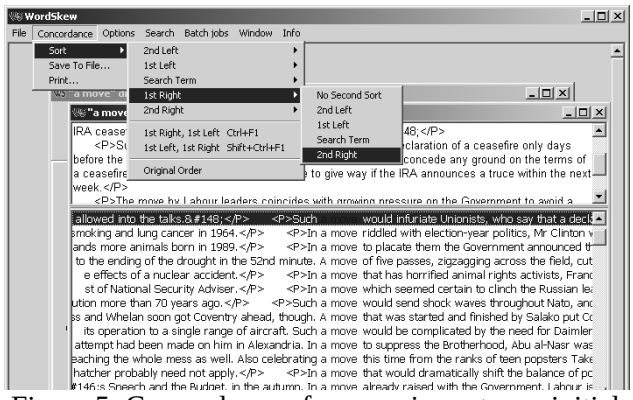

Figure 5: Concordance of *a move* in sentence-initial position

5 Application to contrastive and translation studies

Here the aim is outline the basic features of a text analysis program that relates concordance data to text structure: sentence, paragraphs etc. THe examples presented are not in themselves very interesting, but the intent is to illustrate a tool that allows a finer-grained analysis of the characteristics of translation corpora or of individual translators.

The effect of sentence-splitting on cohesion in German business translations

Mario Bisiada
University of Manchester
mbisiada@fastmail.fm

1 Overview

This paper challenges the commonly held assumption that German is inherently more complex than, for instance, English, by discussing a diachronic increase of sentence splitting in a corpus of German translations of English business and management articles. Contrary to the claim that German translations are invariably more hierarchical because the language favours such a style, translation strategies such as sentence splitting, which lead to paratactic constructions, are commonly observed in the translation corpus and corroborated by a decrease of sentence length in a comparable corpus of non-translated texts. A corpus of unedited draft translations further shows that editor also split sentences regularly, so that the development is not limited to translated languages.

2 Abstract

Sentence splitting has been considered a translation strategy that is enforced by differences in structural norms between the languages involved (Fabricius-Hansen 1996: 558). That assumption seems to be partly based on the fact that research on sentence splitting is thus far largely limited to the German–Norwegian direction of translation (Fabricius-Hansen 1996, 1999; Solfjeld 2008; Ramm 2008). As a ‘high informational density’ language (Fabricius-Hansen 1996), German is said to prefer a hierarchical, hypotactic style, packing ‘much information into each sentence and/or clause by way of a complex syntactic structure’ whereas Norwegian prefers an incremental, paratactic style (Fabricius-Hansen 1996: 558, 1999: 203).

The consensus that seems to emerge from the literature is that translating from a high informational density language to a low informational density language usually favours a translation strategy involving sentence splitting (Solfjeld 2008: 115f). When translating into high informational density languages, on the other hand, ‘structural peculiarities’ such as noun phrase extension and accumulation of adverbial adjuncts are said to ‘allow or even favour hierarchical information packaging to a larger extent than is feasible in English’ (Fabricius-Hansen 1999: 203f).

Translating into a high informational density language such as German, then, should require the opposite strategy to sentence-splitting: 'information collecting [...] and determining which condition on a given discourse referent is to be syntactically downgraded, and how' (Fabricius-Hansen 1996: 561). However, it has not yet been convincingly shown that translators generally introduce cohesion when translating into high informational density languages. Little attention has been paid to sentence splitting in the translation direction English to German, which is what this paper seeks to address.

Using a one million word corpus of English business and management articles and their German translations, the aim of this paper is to test the claim that, as users of a high informational density language, German translators do not (need to) split sentences. Disproving that claim might suggest that sentence splitting is not a strategy that is caused by structural peculiarities of low informational density languages, but rather a feature peculiar to translation in general.

The study finds that, contrary to what seems to be assumed in the literature, German translations exhibit a large amount of sentence splitting, effected both by translators and editors. This is the case especially in more recent translations of 2008 when compared to those from 1982–3, arguing for a shift in the way cohesion is achieved in German business writing. That shift seems to be from hypotactic and paratactic connection on the clause-level to anaphoric pronominal co-reference and sentence-initial conjunctions on the sentence-level.

3 Corpus contents

The study draws on three corpora of business and management articles:

- a translation corpus, which consists of English source texts and their
- published German translations
- a comparable corpus, which consists of German non-translations
- a pre-edited corpus, which consists of English originals, unedited draft translations into German that are yet to undergo editing as well as the versions of these translations that were finally published

The texts in the translation and comparable corpora were published in 1982–3 and 2008, which allows a diachronic analysis of changes in them. The texts in the pre-edited corpus are from 2006–11. The sources for the corpora are the *Harvard Business Review*, an American business magazine, and its licensed German edition, the *Harvard Business Manager*.

References

Fabricius-Hansen, C. 1996. "Informational density: a problem for translation and translation theory". *Linguistics* 34 (3): 521-566.

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

Ramm, W. 2008. „Upgrading of non-restrictive relative clauses in translation: a change in discourse structure?". In C. Fabricius-Hansen and W. Ramm (eds.) *"Subordination" versus "Coordination" in sentence and text: a cross-linguistic perspective*. Amsterdam: John Benjamins.

Solfjeld, K. 2008. "Sentence splitting—and strategies to preserve discourse structure in German–Norwegian translations". In C. Fabricius-Hansen and W. Ramm (eds.) *"Subordination" versus "Coordination" in sentence and text: a cross-linguistic perspective*. Amsterdam: John Benjamins.

Corpus Jerome: issues in the development of a monolingual comparable corpus

Lucie Chlumská
Institute of the Czech National Corpus,
Charles University in Prague
lucie.chlumska@ff.cuni.cz

1 Introduction

The research of the language of translation and its characteristic features has been in the centre of corpus-based translation studies for many years now. To analyze it properly and draw some general conclusions, substantial data resources in the form of various corpora are necessary. Even though there has been a twenty-year-old tradition of corpus compilation in the Czech Republic[1], none of the available corpora was suitable for the research of translated Czech as such. Czech researchers do have a multilingual parallel corpus InterCorp[2] at their disposal, but not a monolingual comparable corpus.

This paper describes the initiative to build a proper comparable corpus of translated and non-translated Czech. It discusses the issues in the development concerning size, source language distribution, genres etc., which are not limited to the Czech situation; they may have implications for other researchers as well.

2 Compilation of the Jerome Corpus

The Jerome Corpus is a monolingual comparable corpus (according to the corpus typology by Laviosa 2002: 36 or Fernandes 2006: 91). It was compiled[3] at the Institute of the Czech National Corpus and made available to public[4] at the end of 2013. It consists of a translational corpus of Czech translations from various languages and a non-translational corpus of Czech originals.

It is a synchronic corpus containing texts published in 1992-2009 (i.e. the modern Czech after the fall of communist regime in 1989). The corpus is lemmatized, morphologically tagged and annotated in terms of standard text information (author, translator, date and place of publication etc.).

[1] The Czech National Corpus (CNC) is one of the largest corpus databases in the world: http://korpus.cz/english/index.php.
[2] Detailed information about the parallel corpus InterCorp available at: http://www.korpus.cz/intercorp/?lang=en.
[3] Within the grant VG027 2013 FA CU.
[4] The corpus can be accessed via KonText interface: http://korpus.cz/english/hledat_v_cnk.php.

3 Main criteria for text selection

Although most comparable corpora used in translation studies do not exceed several million tokens, our objective was to create a very large corpus especially suitable for a quantitative research, i.e. to include as many texts as possible without violating the desired representativeness. This task proved to be almost impossible; it was necessary to make a compromise (see Zanettin 2011: 20), pragmatically sort the objectives according to their importance and then meet the crucial criteria.

With a large size (see table 1) being the most desirable feature, all texts from the CNC database published within the required period were included in the Jerome corpus, provided that:

- They were complete texts (no partial texts or volumes);
- The same author did not have more than three publications in the corpus;
- The same translator did not have more than three translations in the corpus (each one of a different author).

JEROME	Tokens incl. punctuation (TRA/ non-TRA)	Texts
Total	85 065 312	1 526
Fiction	26 551 540 / 26 617 523	394 / 444
Professional	15 949 930 / 15 946 319	382 / 304

Table1: The Jerome Corpus – size and structure

4 Text types and genres

Other important objective was to include more text types[5]: fiction and professional texts. Further division of fiction (such as novels, short stories, poems etc.) had not been taken into account; however, it is included in the text annotation to enable the user to create their own subcorpus.

The CNC texts from the professional domain are further divided into a wide range of genres, such as law, medicine, history, music, chemistry etc. These have been accounted for in a balanced subcorpus (see part 5).

5 Source languages

It is crucial for a translational corpus to be balanced in terms of source languages of translations. However, in Czech, as in many smaller languages, translations from English are three times more

[5] However important, the issue of text types/genres and their definition far exceeds the limited scope of this paper. In this case, the traditional division used in the CNC was used.

common than from any other language. To include the same amount of texts from all available languages would considerably affect the desired corpus size, so a pragmatic approach had to be adopted.

The Jerome Corpus as a whole thus reflects the *reality* of Czech translated literature in the given period[6]; English is by far the prevailing language. However, to make the corpus available for the research of translation universals, a balanced subcorpus was created within the Jerome Corpus. This subcorpus of 5 million tokens includes equal amount of texts translated from 14 typologically different languages in fiction and 6 in professional literature.

6 Additional annotation

To make the corpus as useful and versatile for translation scholars as possible, further information was manually added to the text annotation, such as the author and translator's gender and the year of first edition of the text.

The idea to include information about translator's age (in the form of a year of birth) turned to be impossible to realize (the required data are not available).

7 Conclusion

The development of a comparable corpus has showed that it is necessary to first choose a few main objectives and then try to meet the relevant criteria. The main issue of smaller languages – the source language representation – may be resolved in the form of a balanced subcorpus, whereas a large, yet not balanced corpus may reveal some general characteristics[7] about the translated language as it actually looks like in the eyes of its users.

References

Fernandes, N. 2006. "Corpora in Translation Studies: revisiting Baker's typology". *Fragmentos* 30: 87–95.

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

Zanettin, F. 2011. "Translation and Corpus Design". *SYNAPS - A Journal of Professional Communication* 26: 14-23.

[6]According to the Czech National Library statistics of translated books, available (in Czech) at: http://text.nkp.cz/sluzby/sluzby-pro/sluzby-pro-vydavatele/vykazy.

[7] A case study to support this claim will be included in the presentation.

CIS on screen: a case study on questions and answers

Eugenia Dal Fovo
University of Trieste
eugenia.dalfovo@phd.units.it

This presentation revolves around a Ph.D. research project currently being conducted within the CorIT (Italian Television Interpreting Corpus) project (Straniero Sergio & Falbo 2012) of the Department of Legal, Language, Interpreting and Translation Studies (University of Trieste). The Ph.D. project focuses on the interpreted texts (ITs) of the 2004 American presidential and vice-presidential debates broadcast on Italian television (DEB04 corpus), with the aim of analyzing the question-answer (Q/A) group rendition by interpreters working in an *équipe* in simultaneous mode within the specific constraints of the television setting, *and* without sharing the *hic et nunc* with the primary participants in the interaction (Falbo 2009; Dal Fovo 2012 a, 2012 b).

The presentation starts with an overview of television interpreting as professional activity and research area: capitalizing on early contributions to the literature (*inter al.* Kurz 1985, 2003; Alexieva 1996, 2001; Pöchhacker 1997; Mack 2001; Bros-Brann 2002), the analyst reflects on issues such as constraints and setting-related factors that subsequently led scholars to identify specific norms and strategies, and eventually new quality standards and criteria to be applied to this particular field of reference. The presentation then moves on to provide an illustration of the CorIT corpus and the multiple and unique research opportunities it has provided in the past decade, both for researchers and MA students involved in the project (Dal Fovo 2011). Subsequently, the focus shifts on the main methodological issues that had to be tackled in order to perform a corpus-based analysis in this specific case: indeed, the corpus of analysis, DEB04, serves both as corpus of analysis *per se* and as a "training corpus" (Leech 1997: 9), namely a tool used to try out, select and subsequently 'train' the tagging software (*tagger*) of choice, in order to calibrate it and maximize its rendition when applied to the entire CorIT corpus. Design, collection, transcription and alignment phases will be illustrated.

In the second part of the presentation, data and analysis are presented, with particular attention devoted to the elaboration of the question/answer (Q/A) template of analysis, based mainly on studies on conversation and discourse analysis (Halliday & Hasan 1987; Clark & Brennan 1991; Heritage & Greatbatch 1991; Maley & Fahey 1991; Greatbatch 1992, 1998; Clark 1998; Hale 2001). By means of

conclusion, a broader view of the matter is taken into consideration. Indeed, the analysis raises a series of more general, yet crucial questions regarding communication on television – i.e. the television text, its features and functions – as discourse practice (Straniero Sergio 1999), in which relational aspects and complex participation and organization structures play a major role. Such conditions have significant implications on specific choices and behaviours in terms of discourse and translation attitudes and tendencies (*inter al.* Katan & Straniero Sergio 2003) – either of television interpreters or those taking up their role (e.g. journalists, newscasters, etc.). This generates equally specific users' expectations and more or less prescriptive norms regarding translation in particular and the profession of television interpreters in general (Dal Fovo 2011).

The preliminary outcomes of the present investigation suggest that, despite the considerable amount of research conducted in this field, television interpreting still remains a very elusive subject, whose multi-faceted nature and diverse expressions have yet to be sufficiently identified and defined. As anticipated by Shlesinger (1998) fifteen years ago, and as has already been the case in numerous areas of interpreting studies, the corpus-based approach might prove a decisive tool in order to address and successfully answer some of these questions.

References

Alexieva, B. 1996. Interpreting Mediated TV Events. In Klaudy, K. and Kohn, J. (eds.) *Transferre Necesse Est*. Budapest: Scholastica, 171-174.

Alexieva, B. 2001. Interpreter-Mediated TV Live Interviews. In Gambier, Y. and Gottlieb, H. (eds.) *(Multi)media Translation. Concepts, Practices and Research*. Amsterdam/Philadelphia: John Benjamins, 113-124.

Bros-Brann, E. 2002. Simultaneous interpretation and the media: interpreting live for television. http://aiic.net/page/630/simultaneous-interpretation-and-the-media-interpreting-live-for-television/lang/1 (accessed on 05.01.2014).

Clark, H. and Brennan, S. 1991. "Grounding in communication". In L. Resnick, J. Levine & S. Teasley (eds.), *Perspectives on socially shared cognition*. Washington: Americal Psychological Association, 127-148.

Clark, H.H. 1998. "Responding to indirect speech acts". In A. Kasher (ed.), *Pragmatics: Grammar, psychology and sociology VI*. London/New York: Routledge, 99-147.

Dal Fovo, E. 2011. "Through the CorIT looking glass - and what MA students found there". *The Interpreters' Newsletter* 16, Special Issue on Television Interpreting: 1-20.

Dal Fovo, E. 2012a. "Topical coherence in Television Interpreting: question/answer rendition". In: Straniero Sergio, F. and Falbo, C. (eds.) *Breaking Ground in Corpus-based Interpreting Studies*. Bern: Peter Lang, 187-210.

Dal Fovo, E. 2012b. "Question/answer topical coherence in television interpreting. A corpus-based pilot study". In Kellett Bidoli, C. J. (Ed.) *Interpreting across genres:multiple research perspectives*. Trieste: EUT, 54-77.

Falbo, C. 2009. "Un grand corpus d'interprétation : à la recherche d'une stratégie de classification". In Paissa, P. and Biagini, M. (eds.) *Doctorants et Recherche 2008. La recherche actuelle en linguistique française*, Cahiers de recherche de l'Ecole doctorale en Linguistique française, 3/2009. Brescia: Lampi di Stampa, 105-120.

Greatbatch, D. 1992. "On the management of disagreement between news interviewees". In P. Drew and Heritage, J. (eds.), *Talk at work: interaction in institutional settings*. Cambridge: Cambridge University Press, 268-301.

Greatbatch, D. 1998. "Conversation analysis: neutralism in British news interviews". In A. Bell and Garrett, P. (eds.), *Approaches to Media Discourse*. Oxford: Blackwell, 163-185.

Hale, S. 2001. "How are courtroom questions interpreted? An analysis of Spanish interpreters' practices". In I. Mason (ed.), *Triadic Exchanges*. Manchester: St. Jerome, 21-50.

Halliday, M.A.K. and Hasan, R. 1987. *Cohesion in English*. English Language Series; London/New York: Longman.

Heritage, J. and Greatbatch, D. 1991. "On the institutional character of institutional talk: the case of news interviews". In D. Boden and Zimmerman, D.H. (eds.), *Talk and social structure*. Berkley: University of California Press, 93-137.

Katan, D. and Straniero Sergio, F. 2003. "Submerged ideologies in Media Interpreting". In M. Calzada Perez (eds.), *Apropos of ideology*. Manchester: St. Jerome, 131-144.

Kurz, I. 1985. Zur Rolle des Sprachmittlers im Fernsehen. In Bühler, H. (ed.) *Translators and their position in society*. Xth World Congress of FIT, Proceedings. Vienna: Braumüller, 213-215.

Kurz, I. 2003. Live TV interpreting − A high-wire act?. In Collados Aís, À. and Sabio Pinilla, J. A. (ed.) *Avances en la investigación sobre interpretación*, Granada, Comares, 159-171.

Leech, G. 1997. Introducing corpus annotation. In Garside, R., Leech, G. and McEnery, A. (eds.) *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London/New York: Longman, 1-18.

Mack, G. 2001. Conference Interpreters on the Air - Live Simultaneous Interpreting on Italian Television. In Gambier, Y. and Gottlieb, H. (eds.) *(Multi)Media*

Translation. Concepts, Practices and Research. Amsterdam/Philadelphia: John Benjamins, 125-132.

Maley, Y. and Fahey, R. 1991. "Presenting the evidence: Constructions of reality in court". *International Journal for the Semiotics of Law*, IV (10): 3-17.

Pöchhacker, F. 1997. Clinton speaks German: a case study of live broadcast simultaneous interpreting. In Snell-Hornby, M., Jettmarovà, Z. and Kaindl, K. (eds.) Translation as intercultural communication. Selected papers from the EST Congress, Prague 1995. Amsterdam/Philadelphia: John Benjamins, 207-216.

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

Straniero Sergio, F. 1999. "The interpreter on the (talk) show. Interaction and participation frameworks". *The Translator* 5 (2): 303-326.

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

Using corpora where dictionaries fall short as reference works: A case study of how a Dutch written crime fiction corpus could be used to investigate the occurrence of Dutch *natiolectisms* in crime fiction subtitles

Reglindis De Ridder
Dublin City University
reglindis.deridder2@mail.dcu.ie

This paper analyses the occurrence of Belgian Dutch and Netherlandic Dutch words and expressions, or *natiolectisms* (Martin 2001), in Dutch-language subtitles by using corpus linguistics techniques.

The official Dutch language planning body, *Nederlandse Taalunie*, recognised the two European national varieties of Dutch in 2003 (Nederlandse Taalunie 2003). As is often the case with pluricentric languages (Clyne 1992), one national variety (*natiolect*) is viewed as more prestigious and, in the case of the Netherlands and Belgium, Netherlandic Dutch has generally held sway. In fact, to date, no regional label is used for Netherlandic Dutch words and expressions in the main Dutch dictionary, Van Dale dictionary, and as a result, Belgian Dutch ("Flemish") is, in reality, still described lexicographically as a deviation from "the" Dutch language in this dictionary.

The strong position of the Netherlandic Dutch variety had an impact on both authentic and translated fiction published in the Dutch language area with editors removing Belgian Dutch linguistic features from Belgian Dutch novels prior to publication, and publishing houses mainly hiring Dutch nationals, rather than Belgian nationals, to translate foreign-language fiction into Dutch. Some linguists and translators in the Netherlands and Belgium, however, have started to speak up for a richer, more inclusive written standard. The Dutch, after all, share their language with the majority of the Belgians, and other Dutch-speakers outside of Europe.

Flemish Public Broadcasting (VRT) has always been an important language planner in Dutch-speaking Belgium. In 1998, it officially acknowledged the existence of a Belgian Dutch standard variety and announced that VRT would no longer strictly adhere to the Netherlandic Dutch standard (Hendrickx 1998). Given that VRT's target audience is Belgian, this research investigates if VRT subtitles used in popular crime fiction series have indeed, between 1995 and 2012, increasingly provided a counterbalance to the Dutch publishing industry's traditional approach to edited written texts

by including Belgian Dutch words and expressions in increasing numbers and frequency, rather than replacing such *natiolectisms* by their Netherlandic Dutch counterparts.

However, since Netherlandic Dutch words and expressions are not labelled in Van Dale dictionary, an alternative reference work had to be found to look up the Netherlandic lexical variants that occurred in the subtitles. To this end, a written fiction corpus was built comprising popular crime novels by both Belgian, and Dutch authors published in the same periods the crime series were broadcast. This written crime fiction corpus allowed the frequency of occurrence of lexical variants found in the subtitles to be tested in the linguistic output (i.e. actual language use) of Belgian and Dutch nationals. The assumption is that words and expressions occurring exclusively or predominantly in the Netherlandic Dutch subcorpus and never or hardly ever in the Belgian Dutch subcorpus could be considered Netherlandic Dutch *natiolectisms*.

This research yields interesting data in relation to trends in the use and dissemination of Belgian Dutch variants, on the one hand, and with regard to the use of corpora in diachronic sociolinguistic research, on the other.

References

Clyne, M. 1992. Pluricentric languages: differing norms in different nations. Berlin: Mouton.

Hendrickx, R. 1998. *Het taalcharter* [The language charter] [Online]. Available from: http://www.vrt.be/taal/taalcharter

Martin, W. 2001. *Natiolectismen in het Nederlands en hun lexicografische beschrijving* [Natiolectisms in Dutch and their lexicographic description]. Belgisch Tijdschrift Voor Filologie En Geschiedenis, 79(3), pp.709-736.

Nederlandse Taalunie 2003. *Rapport Variatie in het Nederlands: eenheid in verscheidenheid* [Report on the variation within Dutch: unity in diversity]. [Online]. Available from: http://taalunieversum.org/taalunie/variatie_in_het_nederlands_eenheid_in_verscheidenheid/

Onomatopoeia in Literary Translation: When two languages *bump* into each other

Mohammad Emami
University of St Andrews, UK
me82@st-andrews.ac.uk

Onomatopoeia is a particular use of sound, so that it is no longer an arbitrary part of the linguistic sign but enhances the meaning. As such, it is perhaps more universally exploitable across languages, especially through fiction which can be argued potentially provides more grounds for any author to use onomatopoeia. A particular research into my parallel corpus of 262 American short stories (1,142,943 words) translated into Persian casts light on the relationship between onomatopoeia in English and Persian, how they are transmitted across these languages, and where they appear in a translation without having a counterpart in the source text. In other words, the subject of investigation is how onomatopoeic effects are treated and used by the translators.

An examination of onomatopoeia in the Persian corpus shows that, in the absolute majority of cases, Persian onomatopoeic words appear as the nominal part of compound verbs, hence remain intact, different to their English correspondents which may no longer be distinguished as onomatopoeia, especially when conjugated as a verb. Furthermore, there can be observed a reduplicative structure in Persian words imitating natural sounds. This phenomenon is not heard of in English as an ordinary use of language, nor can it be defined as a requirement of Persian to double every sound heard in the real world. Therefore, this research would also explore at an early stage how, and how systematically, reduplicating words would work in Persian morphology.

One may identify three groups of onomatopoeic words in Persian translations: (1) Persian onomatopoeias corresponding to English onomatopoeias with either similar or dissimilar sounds; (2) Persian onomatopoeias as translations of genuine non-onomatopoeic English words for which either no equivalent is available in Persian or the translator has decided not to use the non-onomatopoeic option; and (3) Persian onomatopoeias as a straightforward option while describing a rather complicated emotion or state. While this simplifies the translator's search for precise equivalents, it also has the potential to improve the fictional reality.

A list of English onomatopoeic words/verbs was created containing all the varieties in which they

may appear in texts, and then looked up in the English corpus to see how many instances of onomatopoeic words exist in various inflectional forms. The resulting concordance was comprised of 3,089 instances in 249 files, meaning that words with a onomatopoeic origin are used in the majority of the short stories in the English corpus, with an average of about 12.4 instances per applicable short story.

The search for Persian onomatopoeia was much more complicated, having no list already available for this language. A series of data extractions and reproductions was therefore designed to build up a list of reduplicative onomatopoeic words used in the Persian corpus. The final list of 118 entries was used to create a concordance which showed the use of onomatopoeia in 151 short stories with a number of 445 instances. The list of reduplicative onomatopoeic words was then used to create a list of single-part words, assuming they may have appeared on their own. The new 30 words were indeed used, though only in 85 instances in 60 short stories, with 12 new short stories in which no reduplicative form was found earlier. In summary, onomatopoeia was used in Persian corpus in 163 short stories, with 530 instances overall, i.e. 3.25 instances per applicable short story.

These findings suggest that, despite onomatopoeic words being more 'visible' in Persian, they are substantially less frequent than in the English corpus in terms of both the number of the short stories they have appeared in, and the frequency of their use in each text. This implies that on many occasions the translators have found themselves sufficiently equipped by non-onomatopoeic Persian words. If the use of onomatopoeia is considered to be in conjunction with the concept of informality, which is the mainstream in fiction writing, it can be said that these translators may have been concerned not to produce translations of over-informality, or perhaps commonplaceness. A supplementary analysis was further undertaken to explore if onomatopoeia is used differently in Persian non-translations, on the short stories written originally in Persian by three of the corpus's translators. Onomatopoeic words were found in these short stories (58 texts of 131,473 words) only in 46 instances in 22 short stories. It can therefore be deduced, on the basis of the current corpus, that using onomatopoeia is not popular amongst Persian writers either, with a provision to investigate this hypothesis in a larger corpus of Persian non-translated literary texts.

References

Halliday, M. A. K. and Matthiessen, C. M., 2004. *An Introduction to Functional Grammar*. 3 ed. London: Arnold.

Hatim, B. and Mason, I., 1990. *Discourse and the Translator*. New York: Longman Group Limited.

Munday, J., 2008. *Introducing Translation Studies, Theories and Applications*. 2 ed. London and New York: Routledge.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J., 1985. *A Comprehensive Grammar of the English Language*. London ; New York: Longman.

Windfuhr, G., 2009. *The Iranian Languages*. London and New York: Routledge.

Phraseological patterns in interpreting and translation: similar or different?

Adriano Ferraresi
University of Bologna
adriano.ferraresi@unibo.it

Maja Miličević
University of Belgrade
m.milicevic@fil.bg.ac.rs

1 Introduction

Research in corpus-based translation and interpreting studies has typically focused on monolingual comparable and/or interlingual parallel comparisons. Recently, intermodal comparisons between translations and interpretations are emerging as a new paradigm in the discipline(s), aiming to shed light on the traits that distinguish one form of language mediation from the other. Previous studies have compared translated and interpreted texts with regard to putative translation universals (Kajzer-Wietrzny 2012: simplification, explicitation and normalization; Bernardini et al. 2012: lexical simplification), as well as distributions of part-of-speech and colloquial terms (Shlesinger and Ordan 2012).

This paper builds on Bernardini et al. (2012), who introduced EPTIC – the *European Parliament Translation and Interpreting Corpus*, a four-way resource composed of simultaneous interpretations paired with their source texts, and the corresponding translations and source texts. Extending the method used by Durrant and Schmidt (2009) to study phraseology in native and non-native English language production, we investigate phraseological patterns in the translated and interpreted Italian components of EPTIC. The method relies on frequency data gathered from an external reference corpus to overcome the data sparseness problem often encountered in studies of translated language (cf. Bernardini 2011).

2 Corpus description

EPTIC builds on the well-known EPIC corpus (*European Parliament Interpreting Corpus*; Sandrelli and Bendazzoli 2005). EPIC's transcripts of interpreted speeches and their source texts were paired with the corresponding translated versions and respective source texts. The language combination represented in the corpus is currently English-Italian, including translations/interpretations in both directions. The corpus is part-of-speech tagged, lemmatised and indexed with the Corpus WorkBench.[8] Each text is aligned (at sentence level) with its source/target and with the corresponding text in the other mode (oral/written).

The corpus contains 392 texts, for a total of about 180,000 words. The bigger, English>Italian portion contains four versions of 81 texts, while the smaller Italian>English portion has four versions of 17 texts. Work is underway to expand the latter segment of the corpus, which in its revised version should have similar sizes across all components.

3 Method

For this study we concentrate on the Italian subcorpora of EPTIC and on two syntactic patterns only, namely *modifier + noun* (e.g. *precedenti osservazioni* 'previous observations') and *noun + modifier* (*comunità internazionale* 'international community').

After extracting the relevant word pairs using part-of-speech information encoded in EPTIC, we gather frequency data about them from itWaC, a large reference corpus of Italian (Baroni et al. 2009). We then classify EPTIC word sequences according to three criteria: frequent vs. infrequent/unattested ($fq \geq 2$ vs. $fq < 2$ in itWaC), and "strong" vs. "weak" collocations based on two lexical association measures, *t*-score ($t \geq 10$ vs. $t < 10$ in itWaC), and Mutual Information (MI≥7 vs. MI<7 in itWaC). *T*-score is expected to highlight "very frequent collocations" (Durrant and Schmidt 2009: 167; e.g. *diritti umani* 'human rights'), and MI to give prominence to "word pairs which may be less common, but whose component words are not often found apart" (ibid.; e.g. *partenariato strategico* 'strategic partnership'). The number of word combinations belonging to infrequent/unattested, high-*t*-score and high-MI sequences is calculated for each text in each subcorpus and expressed as a percentage (e.g. of high-MI combinations out of the total number of word combinations found in a text). Differences in percentages of each type of word combinations in translated and interpreted texts are then tested for significance using Wilcoxon signed rank tests in *R*.[9]

4 Results

The results show that, compared to translations, interpreted texts are characterised by (1) a significantly *higher* percentage of infrequent/unattested word combinations ($V=2229.5$, $p=0.0015$; Figure 1), and (2) a *lower* percentage of high-MI sequences ($V=1017.5$, $p=0.0212$; Figure 2). No statistically significant difference is found in terms of use of high-*t*-score collocations.

[8] http://cwb.sourceforge.net/

[9] http://www.r-project.org/

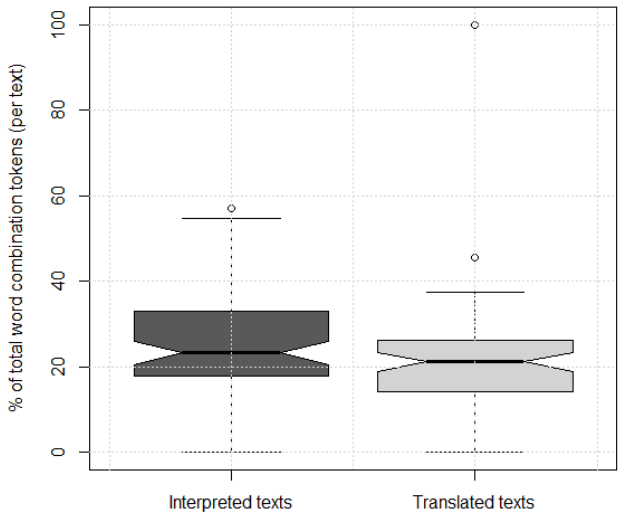

Figure 1. Infrequent/unattested word combinations

/

Figure 2. High-MI word combinations

In other words, interpretations tend to contain more infrequent word combinations and fewer highly idiomatic ones, while being similar to translations when it comes to high-frequency combinations. No difference emerges as significant when the same procedure is applied to the comparable non-mediated written vs. spoken texts (source texts of the Italian>English portion of the corpus), suggesting that the observed features can indeed be seen as specific to translation/interpreting, rather than applying more generally to the distinction between oral and written production.

The paper will conclude by discussing the implications of these results for research on translation/interpreting universals, and highlighting the potential of intermodal corpus resources for corpus-based interpreting/translation studies at large.

References

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. 2009. "The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43 (3): 209-226.

Bernardini S. 2011. "Monolingual comparable corpora and parallel corpora in the search for features of translated language". *Synaps* 26: 2-13.

Bernardini, S., Ferraresi, A. and Miličević, M. 2012. *From EPIC to EPTIC - building and using an intermodal corpus of translated and interpreted texts*. Paper presented at the 46th Annual Meeting of the *Societas Linguistica Europea*, 18-21 September 2013, Split, Croatia.

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

Kajzer-Wietrzny, M. 2012. *Interpreting universals and interpreting style*. Unpublished PhD thesis, Adam Mickiewicz University, Poznań. Available online at https://repozytorium.amu.edu.pl/jspui/bitstream/10593/2425/1/Paca%20doktorska%20Marty%20Kajzer-Wietrzny.pdf.

Sandrelli, A. and Bendazzoli, C. 2005. "Lexical patterns in simultaneous interpreting: a preliminary investigation of EPIC (European Parliament Interpreting Corpus)". *Proceedings from the Corpus Linguistics Conference Series* 1. Available online at http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/ContrastiveCorpusLinguistics/lexicalpatternsinsimultaneousinterpreting.doc.

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

From learner to specialised corpora: Integrating a corpus-based analysis of English NPs in French-English translation teaching

Cécile Frérot **Caroline Rossi**
Univ. Grenoble Alpes, [ILCEA], F
Cecile.Frerot@u-grenoble3.fr Caroline.Rossi@u-grenoble3.fr

1 Introduction

It is generally accepted that terms – especially nouns – are widely used in specialised languages. Besides, complex noun phrases (in which a noun may be modified e.g. by an adjective, another noun or a prepositional phrase) are frequent in scientific texts, especially in medical English, as highlighted by the literature on Terminology and Languages for Specific Purposes (Banks 2001; Depierre 2006; Maniez 2008, 2011; Maniez and Thoiron 2004).
The present study starts from a usage-based, Construction Grammar perspective which articulates both grammar and lexicon (Goldberg 2003): having identified a French construction which is error-prone for students translating into English, we then look at the alternation of two translation equivalents in English corpora.

2 From learner to specialised corpora in translation teaching

In translation teaching, one of the main errors occurring among French students translating into English (L2) is the overuse of the preposition *of* in complex NPs (*quality of the image* vs *image quality*). The overgeneralization of the *[noun] of [noun]* construction could be linked with the prevalence of the corresponding French construction including the preposition *de* (*la qualité de l'image*) which students use as a loan translation. While the varying and contrasted complexity of NPs has long been debated in both reference grammars and translation books (Bouscaren et al. 1992; Vinay and Darbelnet, 2004; Huart and Larreya, 2006), there have been very few corpus-based studies on such thorny issue in translation teaching (Maniez 2011). Overall, corpus-based literature in France is rather poor as far as French-to-English translation is concerned, and only a limited number of French universities have conducted corpus-based studies with the aim of integrating corpus-based data in the classroom to enhance students' translations (Frérot 2013; Kübler 2001; Kübler 2011). The present study suggests some of the contributions corpora can make in a specialised translation environment.

3 A corpus-based study of complex noun phrases in medical English: nominal pre-modification versus prepositional complementation

Our study focuses on the analysis of NPs including the preposition of extracted from a learner corpus on nuclear medicine. The corpus comprises about 5,000 words and includes 17 post-graduate students' English texts. We used AntConc[10] to obtain a list of the most frequent nouns (N1) in the recurring construction (*the*+N1+*of*+N2) -*e.g. the risk of + N2, the response of, the quality of, the choice of, the study of, the position of*. Our premise was that the corresponding Noun+Noun construction may be preferred in at least some of the occurrences (i.e. *treatment choice* vs *the choice of treatment*). In order to verify our assumption, we investigated an English corpus of online articles extracted from ScienceDirect.com and published in *Nuclear Medicine and Biology*. We used AntConc and for each NP identified, we searched the corresponding Noun+Noun construction in order to find which construction prevailed and in which linguistic contexts. For instance, we found a single occurrence of *the risk of cancer* while *cancer risk* has a number of 6 occurrences.

In order to collect more data, we used Scientext - a new, on-line[11] French and English corpus of scientific texts, which includes 13 million words of research articles in English (from the fields of medicine and biology). We found 21 occurrences of *the risk of cancer* while *cancer risk* has a total number of 774 occurrences. A closer look at the data shows that *cancer risk* is often nested within longer terms such as *lung cancer risk*, or *breast cancer risk*, while the occurrences of *the risk of cancer* are found in more abstract contexts, e.g. complementing verbs such as *to increase*.

4 Pedagogical applications and future work

Working with students specialising in translation, we intend *(i)* to raise student awareness of how valuable authentic texts can be in translation (Zanettin 2002; Bernardini and Castagnoli 2008) and *(ii)* to help students provide more accurate and idiomatic translations of complex NPs. To this end, starting from students' errors in our learner corpus and then having our students explore to what degree a given construction is best suited by searching specialised

[10] A freeware concordance program available at http://www.antlab.sci.waseda.ac.jp/software.html
[11] Available at http://scientext.msh-alpes.fr/

corpora may prove motivating for students and relevant for their translations.

This perspective does not only involve using corpora and concordancers directly in the classroom in the vein of corpus-based activities designed elsewhere (Frérot 2009); it also aims at providing students with lists of bilingual NPs as well as frequency and contextual data. The data may also be used to create cloze activities and tests in order to best address this error-prone and scarcely debated translation issue from a corpus-based perspective.

References

Banks, D., (éd.) 2001. *Le groupe nominal dans le texte spécialisé*. Paris, L'Harmattan, pp. 117-136.

Bernardini, S. and Castagnoli, S. 2008. Designing a Corpus-based Translation Course for Translation Teaching and Translator Training. *International Journal of Translation*, vol. 21, n°1-2, pp. 133-147.

Bouscaren, J., Chuquet, J., Danon- Boileau and L., Flintham, R. 1992. *Introduction to a linguistic grammar of English : an utterer-centered approach*, Paris, Ophrys.

Depierre, A. 2006. De l'utilisation de textes spécialisés pour l'enseignement de la terminologie dans le domaine de l'anglais médical, in *Applications et implications en sciences du langage*, dir. Léglise Isabelle, Emmanuelle Canut, Isabel Desmet et Nathalie Garric. Paris : L'Harmattan, pp. 257-268.

Frérot, C. 2009. Designing a Corpus-based Translation Course for Translation Teaching and Translator Training. *International Journal of Translation*, vol. 21, n°1-2, pp. 133-147.

Frérot, C. 2013. Incorporating Translation Technology in the Classroom: Some Benefits and Issues on Exploiting Corpora and Corpus-Based Translation Tools. *Selected papers from the EST Congress, Leuven 2010*. Catherine Way, Sonia Vandepitte, Reine Meylaerts and Magdalena Bartłomiejczyk (eds.). Amsterdam: Benjamins Translation Library, vol. 108, pp. 143-166.

Goldberg, A. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Science* 7(5), pp. 219–24.

Huart, R. and Larreya, P. 2006. *Les constructions Nom+Nom*, collection « gramvoc », Ophrys

Kübler, N. 2001. Corpora in Terminology and Translation teaching: methodological approach. In: de Cock, S. G. Gilquin, S. Granger, and S. Petch-Tyson (eds), *Proceedings of the ICAME 01 Conference*. 2001, pp. 53-55.

Kübler, N. 2011. Working with different corpora in translation teaching. In Ana Frankenberg-Garcia, Lynne Flowerdew, and Guy Aston (eds) *New Trends in Corpora and Language Learning*. London: Continuum, pp. 62-80.

Maniez, F. 2008. Using the Web and corpora as language resources for the translation of complex noun phrases in medical research articles. *Panacea*, n° 26.

Maniez, F. 2011. L'apport des corpus spécialisés en terminographie multilingue : le cas des groupes nominaux de type Nom-Adjectif dans la langue médicale. *Meta*, vol. 56, n°2 : 391-406.

Maniez, F. and Thoiron, P. 2004. Les groupes nominaux complexes dans le décodage et la traduction en langue de spécialité: quelles ressources lexicales pour l'apprenant en anglais médical ? In T. Lino (ed.): *Vocabulaire de spécialité et lexicographie d'apprentissage en langues-cultures étrangères et maternelles*. Paris: Didier Érudition, 327-346.

Vinay, J.-P. and Darbelnet, J. 2004. *Stylistique comparée du français et de l'anglais*. Didier Edition.

Zanettin, F. 2002. Designing a Corpus-based Translation Course for Translation Teaching and Translator Training. *International Journal of Translation*, vol. 21, n°1-2, pp. 133-147.

Using COMENEGO for specialised phraseographic purposes in Spanish and French

Daniel Gallego-Hernández
University of Alicante
daniel.gallego@ua.es

1 Introduction

Problems arising from phraseology in specialised translation are often related to target language production. This is also the case for business translation, whose clients expect from translators "la connaissance du jargon du secteur" and expect that translation "soit dans le ton" (Durban 2005: 66).

Unfortunately, there are currently few phraseographic products that can be used as resources in business translation from French into Spanish and vice versa. This might be due to a terminological problem stemming from a certain degree of vagueness added to this kind of units. For instance, Aguado de Cea (2007: 56-58) identifies different labels (*multi-word terminological phrases*, *phraseology*, *terminological phrasemes*, *specialized lexical combinations*, *collocations*) which may involve various concepts such as multiword terms including a nominal element, unambiguous formulaic expressions, lexical combinations which include both a verb and a term or even lexical expressions which belong to a single specialised field.

Against that backdrop, the COMENEGO (Corpus Multilingüe de Economía y Negocios) project may contribute to compensate the lack of phraseology-related products in French and Spanish. The main aim of this project is to create a stable electronic corpus which can be used by translation practitioners (professionals, trainees and trainers). COMENEGO is also a comparable pilot corpus which has around 19 million words (the Spanish corpus has around nine million words and the French one has also around nine million words) (Gallego-Hernández & Krishnamurthy 2013). As for the French component of the project, we are currently carrying out different surveys on professional translators and clients in order to both justify the choice of topics and genres in French-Spanish and Spanish-French translation and to analyse the uptake of corpora among translators and describe their use of this kind of translation resource (Gallego-Hernández forthcoming).

Once the most common topics and genres are identified in different languages, the project will be able to proceed to reclassify or add new texts to the corpora already compiled, and to start a new stage related to the extraction of terminology and specialised phraseology from different domains and textual genres of COMENEGO.

2 Methodology

This contribution is directly related to this last issue: phraseology extraction. In particular, we deal with collocations (Sinclair 1991). We illustrate how to extract lexical collocations (Benson et al. 1986) containing a terminological node and different verbs. For instance, the term *capital*, which is one of the most frequent terms in both corpora (French and Spanish), may be initially exploited with Antconc's collocates function. Tables 1 and 2 show verbs that collocate with *capital* in Spanish and French:

Rank	Freq	Freq(L)	Freq(R)	Collocate
28	126	0	126	suscrito
29	110	2	108	asegurado
32	87	0	87	invertido
34	85	0	85	garantizado
46	55	0	55	circulante

Table1: Spanish collocates of *capital*

Rank	Freq	Freq(L)	Freq(R)	Collocate
21	154	1	153	garanti
27	106	5	101	souscrit
29	94	0	94	restant
31	88	0	88	investi
44	53	0	53	versé

Table2: French collocates of *capital*

These first results not only show some coincidences such as *capital suscrito/capital souscrit*, *capital invertido/capital investi* or *capital garantizado/capital garanti*, but also imply a starting point of research which can be complemented with Antconc's concordance tool.

In this sense, we can enter expressions such as garant*@@capital++|capital++@@garant* or garant*@@capit++|capit++@@garant* in order to explore such units in greater depth:

```
… tiene garantizado el capital invert…
…ión que garantiza un capital en una …
…con el capital 100 % garantizado que…
… que le garantiza el capital aportad…
…2 años y garantía de capital nominal…
…futuro garantizando un capital para …
…ntant du capital garanti (avant 65 a…
…traite à capital 100% garanti. En sa…
…ent. Une garantie du capital et de l…
…is d'une garantie en capital. Les pr…
… € de capitaux mobiliers garantis  P…
…ité et garantie du capital à l'échéa…
…ntant du capital à garantir (par exe…
```

This selection of concordances shows, among other things, similarities between the two languages but also how the node *capital* works with the verbs *garantizar* and *garantir* and their deverbal categories.

3 Results

Future results of this stage may be used not only to create phraseological glossaries or dictionaries for business translators but also to provide an empirical basis which may help us to objectively classify the texts in the corpus.

References

Aguado de Cea, G. 2007. "La fraseología en las lenguas especializadas", In E. Alcaraz Varó et al. (eds.) *Las lenguas profesionales y académicas*. Madrid: Ariel.

Benson, M. Benson, E. and Ilson, R. 1986. *The BBI combinatory dictionary of English. A guide to word combinations.* Amsterdam/Philadelphia: John Benjamins.

Durban, C. 2005. "La traduction financière. Tendances pour l'avenir", In D. Gouadec (ed.) *Traduction, Terminologie, Rédaction. Actes des universités d'été et d'automne et du colloque international Traduction spécialisée chemins parcourus et autoroutes à venir traduire pour le web.* Paris: La maison du dictionnaire.

Gallego-Hernández, D and Krishnamurthy, R. 2013. "COMENEGO (Corpus Multilingüe de Economía y Negocios): design, creation and applications". *Empirical Language Research Journal* 8.

Gallego-Hernández, D. forthcoming. "The use of corpora as translation resources. A study based on a survey of professional translators". *Perspectives. Studies in Translatology.*

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: University Press.

COMENEGO: Contrasting language varieties in three languages

Daniel Gallego-Hernández
University of Alicante
daniel.gallego@ua.es

Patricia Rodríguez-Inés
Autonomous University of Barcelona
patricia.rodriguez@uab.es

1 Introduction

Although there are various specialised electronic corpora for the business and/or economics field in existence, the majority of them only include texts in English, do not include a wide range of textual genres, are nowadays obsolete or too small, and are certainly not translator-oriented. There was thus a definite need to compile a large, up-to-date, freely accessible, multilingual, multigenre corpus of business and economics texts.

Through the corpus COMENEGO (Corpus Multilingüe de Economía y Negocios) we are seeking to fulfil the need in question. The corpus, which has solid foundations and is progressing steadily, includes Spanish and French components (Gallego-Hernández & Krishnamurthy 2013) and a new English component (Rodríguez-Inés forthcoming), all of which are currently undergoing analysis and expansion.

The main difference among them is related to the time when these corpora were compiled: the Spanish and French pilot corpora were built using external criteria (pragmatic parameters mainly based on Cassany 2004) and intuitive judgments. Furthermore, the results of a survey on the practice of translation in the field of business and economics from French into Spanish and vice versa to justify the choice of topics and genres (Gallego-Hernández 2013a) were not available yet at the time when the two corpora were being compiled. In contrast, the English pilot corpus was built after having the results of the survey (Tolosa-Igualada forthcoming) and having started analysing the initial categories of the French and Spanish corpus in order to justify the text classification into seven discursive categories (commercial, didactic, legal, organizational, press, scientific, technical) which were initially arrived at. (Krishnamurthy & Gallego-Hernández 2012; Gallego-Hernández 2013b; Suau-Jiménez forthcoming).

In this presentation, first we will briefly discuss the characteristics of the three pilot corpora and compare the stages involved in the building of the

corpora. Then we will focus on one of the stages involved in the COMENEGO Project: comparing the seven discursive categories.

2 Methodology

In order to try to answer the question of whether these categories have any internal linguistic features that support/confirm their taxonomic validity, we will use corpus linguistics tools that allow us to obtain various kinds of analytical output from the three corpora: Antconc's word frequency lists, concordances, and n-grams. We will compare potential category-specific content words and previous French and Spanish metadiscursive analysis based on Hyland (2005), with new results related to the English corpus.

3 Results

The results of the analysis of the three pilot corpora should help us to identify imbalances and deficiencies which should be addressed, and also to confirm or reject the classification of the corpus texts so that it can be implemented in the virtual platform which is still under construction and will allow users to exploit the corpus.

References

Cassany, D. 2004. "Explorando los discursos de las organizaciones". In A. van Hooft Comajuncosas (ed.). *Textos y discursos de especialidad. El español de los negocios.* Amsterdam/New York: Rodopi.

Gallego-Hernández, D and Krishnamurthy, R. 2013. "COMENEGO (Corpus Multilingüe de Economía y Negocios): design, creation and applications". *Empirical Language Research Journal* 8.

Gallego-Hernández, D. 2013a. "Que traduisent les traducteurs économiques du français vers l'espagnol et de l'espagnol vers le français? Étude basée sur une enquête". Unpublished work.

Gallego-Hernández, D. 2013b. "A Comparative Corpus-Based Analysis of Metadiscourse in COMENEGO". Paper presented at *ICLC 7 - UCCTS 3,* Gent Universiteit.

Krishnamurthy, R. and Gallego-Hernández, D. (2012): "Discursive analysis of textual resources of COMENEGO". Paper presented at *IV Congreso Internacional de Lingüística de Corpus CILC2012,* University of Jaén.

Rodríguez-Inés, P. forthcoming. "COMENEGO: Compilación del corpus piloto en inglés y primeros análisis". *VERTERE. Monográficos de la Revista Hermēneus.*

Suau-Jiménez, F. forthcoming. "*Preparación pre-traductológica de textos de economía a través del análisis metadiscursivo interpersonal: estudio piloto del corpus COMENEGO". VERTERE. Monográficos de la Revista Hermēneus.*

Tolosa-Igualada, M. forthcoming. "Dime qué traduces y «les» diré quién eres. Estudio basado en encuestas acerca de los documentos traducidos por traductores económicos (inglés-español y español-inglés)". *VERTERE. Monográficos de la Revista Hermēneus.*

The translation of source language lacunas: An empirical study of the Over-Representation Of Target Language Specific Features and the Unique Items hypotheses

Lidun Hareide
University of Bergen
Lidun.hareide@if.uib.no

1 Introduction

The aim of this paper is to empirically test two hypotheses posited on the proposed translation universal over- or under-representation of target-language specific features, these being the Overrepresentation of Target-Language Specific Features Hypothesis (Baker 1993, 1995, 1996) and the Unique Items Hypothesis (Tirkkonen-Condit 2001, 2004). Although mutually exclusive, both Baker's and Tirkkonen-Condit's hypotheses have been attested by empirical research. The hypotheses are tested on the language pair Norwegian-Spanish, using the Spanish gerund as a test object. In order to realize this project, the 4.1 million word Norwegian-Spanish Parallel Corpus (NSPC) was compiled (Hareide and Hofland 2012). The Spanish Corpus de Referencia de Español Actual [12] was used as a reference corpus.

2 Theoretical background

Research on the six hypotheses collectively known as the Translation Universals Hypothesis [13] (Baker 1993) constitutes one of the main branches ofempirical Translation Studies. One of the most controversial and most interesting of these from a research perspective is the hypothesis that one can observe "a general tendency to exaggerate features of the target language" (Baker 1993: 244). This hypothesis is further developed in Baker (1999: 183), and was put forward on the basis of earlier research by several prominent scholars such as Toury (1980) and Vanderauwera (1985). Vanderauwera suggests that translations "over-represent features of their host environment in order to make up for the fact that they were not meant to function in that environment" (Baker 1993: 245). Empirical research by Halverson (2007) also supports this hypothesis.

Sonja Tirkkonen-Condit argues against Baker's hypothesis of over-representation of features of the target language (Tirkkonen-Condit 2004: 177). Tirkkonen-Condit proposes the Unique Items Hypothesis, where she argues that these structures are in fact under-represented in translations, because there are no corresponding structures in the source language that will trigger their use. In her opinion; "Since they are not similarly manifested in the source language, it is to be expected that they do not readily suggest themselves as translation equivalents, as there is no obvious linguistic stimulus for them in the source text" (Tirkkonen-Condit 2004: 177). (For a discussion of the Unique Items Hypothesis, see (Chesterman 2007)). This hypothesis is supported by empirical research by Kujamäki (2004), Eskola (2004), (Rabadán, Labrador, and Ramón 2009) Vilinsky (2012), and (Capelle 2012).

3 Methodology

In order to empirically test the two hypotheses, I had to establish empirically that the Spanish gerund in fact does constitute a unique item in translations from Norwegian. This was done by analyzing the structures in the source language Norwegian that gave rise to the Spanish gerund in translations. From each of the texts in the NSPC, a random sample of 20% of the sentences containing Spanish gerunds and their corresponding source-language sentences was extracted (a total of 1597), and the structures in the source language that triggered the use of the Spanish gerunds were established. In order for the Spanish gerund to qualify as a unique item in translations from Norwegian, this study would have to establish that no single Norwegian structure triggers the use of the Spanish gerund. Instead a wide variety of structures would give rise to the Spanish gerund in translations from Norwegian.

In addition, the number of gerunds in the NSPC and in a subcorpus extracted from the CREA that corresponds to the NSPC with regard to sampling frame (time-span, Spanish variety and genres) was established in order to calculate the frequency of the Spanish gerund in the two corpora using the log-likelihood statistical measure.

4 Results

The Spanish gerund was found to be a unique item in translations from Norwegian as a total of 14 structures ranging from finite verbs to prepositions and prefixes were found to be the source-language triggers of the Spanish gerund. Finite verbs and aspectual structures (structures that perform similar functions as those expressed by aspect in other languages) (Faarlund, Lie, and Vannebo 1997: 644 -

[12] REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. Corpus de referencia del español actual. <http://www.rae.es>

[13] This hypothesis is also referred to as the Features of Translation Hypothesis.

5) were found to be the most frequent.

Even so, the Unique Items Hypothesis was refuted in my study, as the Spanish gerund was found to be significantly over-represented in Spanish translated from Norwegian. Consequently the Over-representation of Target-Language Specific Features received support.

5 Concluding remarks

The fact that the Unique Items Hypothesis is refuted in this analysis raises an intriguing question: What is needed for the Unique Items Hypothesis to receive support? One suggestion might be that the Unique Items Hypothesis requires a language pair composed of languages that are very typologically different, such as Finnish (an Uralic language) in contrast to Indo-European languages. Most studies on the Unique Items Hypothesis, such as Tirkkonen-Condit (Tirkkonen-Condit 2001, 2004), Kujamäki (Kujamäki 2004), and Eskola (Eskola 2004), have been conducted on data from the Corpus of Translated Finnish (CTF). However, recent research by Vilinsky (Vilinsky 2012) and Capelle (Capelle 2012) provide support for the hypothesis using the language pairs English-Spanish and French-English respectively, indicating that factors other than typological difference may enter into the equation.

References

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies." In *Text and Technology: in honour of John Sinclair*, edited by Gill Francis, Tognini-Bonelli,E. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Baker, Mona. 1995. "Corpora In Translation Studies: An Overview and Some Suggestions for Future Research." *Target* no. 12:241-266.

Baker, Mona. 1996. "Corpus-based translation studies: The challenges that lie ahead." In *Terminology, LSP and translation: Studies in language engineering*, edited by Harold L Somers, 175-186. Amsterdam: John Benjamins.

Baker, Mona. 1999. "The role of corpora in investigating the linguistic behaviour of professional translators." *International Journal of Corpus Linguistics* no. 4 (2):1-18.

Capelle, Bert. 2012. "English is less rich in manner-of-motion verbs when translated from French." *Across Languages and Cultures* no. 13 (2):173-195.

Chesterman, Andrew. 2007. "What is a unique item?" In *Doubts and Directions in Translation Studies*, edited by Yves Gambier, Miriam Shlesinger and Radegundis Stoltze. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Eskola, Sari. 2004. "Untypical frequencies in Translated language. A Corpus based study on a literary corpus of translated and non-translated Finnish " In *Translation Universals, Do They Exist?*, edited by Anna Mauranen and Pekka Kujamaki. Amsterdam/Philadelphia: John Benjamins.

Faarlund, Jan Terje, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Halverson, Sandra. 2007. "Investigating Gravitational Pull in Translation: The Case of the English Progressive Construction." In *Text, Processes, and Corpora: Research Inspired by Sonja Tirkkonen-Condit*, edited by Riita Jääskeläinen, Tiina Puurtinen and Hilkka Stotesbury. Savonlinna: Savonlinna School of Translation Studies 5.

Hareide, Lidun, and Knut Hofland. 2012. "Compiling a Norwegian-Spanish Parallel Corpus: methods and challenges." In *Quantitative Methods in Corpus Based Translation Studies*, edited by Michael Oakes and Meng Ji. Amsterdam: John Benjamins Publishing Company.

Kujamäki, Pekka. 2004. "What happens to 'unique items' in learners' translation." In *Translation Universals: Do they exist?*, edited by Anna Mauranen & Pekka Kujamäki. Amsterdam/Philadelphia: John Benjamins.

Rabadán, Rosa, Belén Labrador, and Noelia Ramón. 2009. "Corpus-based contrastive analysis and translation universals. A tool for translation quality assessment English -> Spanish." *Babel* no. 55 (4):303-328.

Tirkkonen-Condit, Sonja. 2001. Unique items - over- or under-represented in translated language? In *The Third International EST Congress*. Copenhagen, Denmark.

Tirkkonen-Condit, Sonja. 2004. "Unique Items - over - or under-represented in translated language?" In *Translation Universals - Do they exist?*, edited by Anna Mauranen and Pekka Kujamäki, 177-184. Amsterdam/Philadelphia: John Benjamins.

Toury, Gideon. 1980. *In Search of a Theory of Translation*. Tel Aviv: Porter Institute.

Vanderauwera, R. . 1985. *Dutch Novels Translated into English: The Transformation of a "Minority" Literature*. Amsterdam: Rodopi.

Vilinsky, Bárbara Martínez. 2012. "On the lower frequency of occurrence of Spanish verbal periphrases in translated texts as evidence for the Unique items hypothesis." *Across Languages and Cultures* no. 13 (2):197-210.

General tendencies and variations of translational English across registers

Xianyao Hu
Southwest University, China
huxyao@hotmail.com

Richard Xiao
Lancaster University
r.xiao@lancaster.ac.uk

Translation Universals and its hypotheses, despite the challenges or oppugnations since its birth in early 1990s, has become a meaningful and valuable concept in Translation Studies in that it gives rise to the idea that the translated texts may be a special and distinctive variant of language, or the "third code" (Frawley 1984). These distinctive features could be the result of the interaction of the cognitive process, socio-cultural context, and language transfer that translating involves, and hence a key to unveiling the fundamental factors of translating. Empirical studies of the TUs by far have shown a rudimentary picture of these universals or general tendencies of translational language: on the one hand, the translational language tends to conform to the TL norms through simplification and explicitation in order to increase its acceptability to the target language community; on the other, the translational language also shows tendencies of breaking the TL norms, changing the meaning of words and expressions, and using creative collocations to experiment on the TL potentials.

Given all contributions and potentials of TUs studies, it is clear that these studies still need to address many of the theoretical challenges and methodological problems. Seen merely from the empirical and methodological side, TUs studies have been mostly limited in a small number of typologically close-related European languages, particularly the translation of these languages to/from English. There is a perceivable lack of TUs studies between genetically distant languages, e.g. English and Chinese. When describing specific linguistic and textual features of translational language, researchers tend to talk about translation universals from different perspectives, using examples in different languages, and focusing on particular linguistic levels, as a result, contradictory evidence of the same universal hypothesis was presented at different linguistic levels in different languages. Even in the same target language, the common features of the translational variant are often too diversified to form a consistent whole picture of the translational language in question. For example, the research of translational English since the 1990s' was mostly based on the *Translational English Corpus* (TEC) built by Mona Baker. As the first corpus of translational English, TEC laid the ground for the later corpus-based translated studies, however, TEC is not a balanced corpus, containing only four types of texts, with fiction as the majority. It is, accordingly, not comparable to most of the current balanced English corpora either in terms of sampling or structure. Consequently, it is impossible to use TEC to study the variations of the general tendencies/features across genres, while the latter is the key to tackling the contradictory evidence in TUs studies.

The current paper is a part of the ESRC and RGC HK joint research project, "comparable and Parallel Corpus Approaches to the Third Code: English and Chinese Perspectives". The main aim is to identify common features of translated English texts and to investigate variations in such features across registers/genres based on the balanced corpora. But the inter-lingual comparison is not possible until the comprehensive and systematic studies of translational English and translational Chinese are done respectively. For this purpose, a balanced *Corpus of Translational English*（COTE）was built by Richard Xiao at Lancaster University. The COTE corpus is a one-million-word balanced comparable corpus of translated English designed as a translational counterpart of the *Freiburg–LOB Corpus of British English* (F-LOB). It is intended to match the F-LOB corpus as closely as possible in size and composition, but is supposed to represent English in translation of the early 1990s. Similar to the 'Brown family' corpora, COTE contains 500 texts of around 2000 words each, distributed across 15 text categories. This paper will first of all investigate the general tendencies of translational English (COTE) in contrast to non-translational or native English (F-LOB). Starting from the macro statistical analysis of the corpora, we will present general features of translational English, ranging from lexical density, wordlist analysis, distribution of word classes to mean word/sentence length, etc. With these general tendencies, we hope to set the scene for the detailed discussion of the TUs, i.e., simplification, explicitation, normalization, convergence, under-representation and SL shining through in translational English. Due to the balanced structure, well-sampled repetitiveness and increased comparability to the native English corpus (F-LOB), the statistical analysis of this research will be more trustworthy and comprehensive. And more importantly, we will be able to look into the variations of these features across genres (for example, news, general prose, academic and fiction) in order to reach at a more fine-grained view of translational English.

References

Baker, M. 1993. "Corpus linguistics and Translation Studies: Implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honor of John Sinclair*, 233-250. Amsterdam: John Benjamins.

Frawley, W. 1984. "Prolegomenon to a theory of translation", in W. Frawley (ed.) *Translation: Literary, Linguistic and Philosophical Perspectives*, 159-175. London: Associated University Press.

Laviosa, S. 2002. *Corpus-Based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.

Kenny, D. 2001. *Lexis and Creativity in Translation. A Corpus-Based Study*. Manchester: St. Jerome Publishing.

Russian Learner Translator Corpus in translator training

Tatyana Ilyushchenya
Tyumen State University
tatyana1223@mail.ru

Marina Kovyazina
Tyumen State University
Makovyazina@mail.ru

Maria Kunilovskaya
Tyumen State University
mkunilovskaya@gmail.com

1 RusLTC as the source of data

The purpose of this proposal is to develop a series of interactive on-line exercises for Russian translator trainees translating out of English to prevent most typical translation errors. The research is based on the Russian Learner Translator Corpus (Kutuzov et al. 2012) which is being developed as a joint project of translator trainers from the Tyumen State University and computational linguists from the Higher School of Economics (Russia).

RusLTC is a parallel corpus of translation trainees' target texts aligned with their sources in English and Russian, which are translators' working languages regardless of the direction of translation. Learner translators' mother tongue is Russian. The project sets out to create an available and reliable resource to be used in translation studies research and to inform translation pedagogy.

As of December 2013 the Corpus size about 1 mln tokens split almost equally among English and Russian texts regardless of whether source or target. The Corpus includes over 200 English sources and approx. 900 Russian translations, and over 30 Russian sources and approx. 600 English translations, the explanation for the discrepancy in the figures being that that the Corpus contains multiple translations of the same source. The number of translations varies from 1 to more than 60.

All translations are done by translator trainees or non-professional translators at 10 Russian partner universities under different conditions – as routine home assignments, as test classroom translations, as part of translation contest programmes. The relevant information about those conditions and affiliations (when available) is included in meta data searchable via the Corpus interface. The query tool supports lexical search for both sources and targets and returns all occurrences of the query item in respective texts along with their targets/sources aligned at sentence level. A new release of the query

tool supports lemmatization and POS-tagging and is currently in alpha testing). While running such queries it is possible to narrow them down by specifying particular conditions of translations, types of trainees or source text genre. There is an option to view full texts and corresponding meta data.

The current research is based on a small translation error-tagged subcorpus which includes about 200 manually error-tagged translations, mostly into Russian. We have used RusLTC server installation of **brat**, a program for text annotation (Stenetorp et al. 2012) to create error annotations. It operates on the error typology designed for this purpose.

2 Inter-rater reliability of data

To ensure the reliability of the manual mark-up we have carried out three inter-rater agreement tests, which 1) showed a greater degree of consensus between raters applying error-based approach to the quality of translation in comparison with rating based on holistic evaluation of translations; 2) proved that the raters mostly agree when adding annotations to more critical content errors, while tend to differ in opinions when judging about less significant language errors; 3) indicated that the inter-rater agreement is higher for poor translations than for good ones; 4) showed that additional training for raters and improvements introduced into the classification between the consecutive experiments did increase the reliability of the error-tagging. The agreement between three raters, who evaluated 22 translations of one text, reached the acceptable degree, expressed as Krippendorff's Alpha coefficient of α=0,734.

3 Class-room use of translation error-tagged subcorpus

The current routine use of error-annotated translations consists in 1) discussion and analysis of most common and individual mistakes marked by the teacher; 2) blind annotating mistakes in peer translations and explaining them (including in Notes to each tag); 3) editing tagged translations (both one's own and peers'); 4) comparing translations of the same text and explaining the advantages and disadvantages of the offered variants. Most of these activities focus attention on the post-translation stage of self-reviewing which is important to produce quality translation and is often overlooked during training. Another way in which we utilize previous translations of the same source is getting students to look at somebody else's mistakes before translation to highlight potentially dangerous phrases and increase awareness of possible problems.

Apart from that we propose to use the marked-up translations to identify statistically most common translation-induced mistakes in English-Russian translations and develop excises to prevent them. We will use them in translator training and compare the results of the entry and final tests as to the quality improvements in the targeted area. If successful, we plan to create a corpus-driven e-learning course that will address most frequent mistakes taking into account the description of best practices offered by MeLLANGE consortium (2007).

References

MeLLANGE (Multilingual eLearning in LANGuage Engineering). 2007. *Best practices in e-learning content creation and development*. Available at http://mellange.eila.univ-paris-diderot.fr/Best_practices.pdf

Kutuzov, A.B., Kunilovskaya, M.A., Oschepkov, A.Y, Chepurkova, AY. 2012. "Russian Learner Parallel Corpus as a Tool for Translation Studies". In Proceedings of the *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*. Issue 11. Vol. 1 of 2: 362-369. Available at http://www.rus-ltc.org/references/dialog.pdf

Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S. and Tsujii J. "brat: a Web-based Tool for NLP-Assisted Text Annotation". In Proceedings of the *Demonstrations Session at EACL 2012*. Available at http://brat.nlplab.org/index.html

Parallel corpus and metatext

Anna Kisiel
Institute of Slavic Studies,
Polish Academy of Science
ania-kisiel@tlen.pl

1 Introduction

The paper consists of two parts. Part one presents methods of applying corpus linguistics to semantic analysis of metatext in two – and more – languages. Part two shows how the outcome of such an analysis can be used in multilingual lexicography.

2 Database for the project

A thorough contrastive study of metatext in Slavic languages has not been yet provided. Both dictionary entries and few available linguistic descriptions limit the picture to suggesting the best equivalents. It is highly unlikely that in two languages there are metatextual language units bringing exactly the same meaning, even when languages in question belong to the same language group. There is, on the other hand, a possibility that the same meaning is expressed differently in two languages - by a language unit in one of them and syntactic construction in the other.

At present, a big parallel Russian-Polish-Bulgarian corpus is being constructed as a part of Clarin project (by Semantics and Corpus Linguistics Team in Institute of Slavic Studies, Polish Academy of Science). The corpus gathers linguistic data allowing comparison of ways of expressing the same meaning in three languages representing three different branches of Slavic languages.

3 Part one. Generalising particles and the problem of equivalence

The first, introductory, part of the paper is of semantic character and presents different contexts containing Polish generalising particles (*ogólnie*, *w ogólności*, *generalnie* ≈ *in general*, *generally*) as well as two language units based on the same root but not belonging to the group (*ogółem*, *w ogóle* ≈ *altogether*, *on the whole*) and ways of expressing the same meaning in Russian and Bulgarian. Such a comparison is here given to show the following problems:

a) are Bulgarian *генерално*, *като цяло*, *обикновено*, *общо полностью*, *целиком*, *по большому счёту* real language units or still constructions?,

b) do all these language units represent metatextual level?,

c) if so, are they semantically identical with Polish generalising particles?,

d) how to approach those of Bulgarian and Russian contexts in which a Polish generalising particle has no visible translation?

and finally

e) what to do in a situation when one language unit is seen as an equivalent of different language units in another language, units that do not share semantic compounds, for example Russian *в целом* as equivalent of Polish *generalnie*, *ogólnie* and *w ogóle* or Bulgarian *общо* as equivalent of Polish *w ogólności* i *ogółem*.

The situation described in d) requires a particularly detailed analysis of a text. It is hardly probable that a metatextual comment was ignored by a translator. More likely, the meaning carried by the comment is hidden in a text preceding or – less likely – following the equivalent sentence. Such situations (characteristic for *w ogólności*) offer an insight into how a language manages to express a metatextual meaning unfamiliar to its system.

4 Part two. Parallel corpus as a tool in language teaching

The second part shows – basing on what has previously been said – how to make multilingual dictionary entries more useful. Such correction is necessary since:

1. Most multilingual dictionaries make a mistake of giving a sequence of equivalents, forcing a user to choose the most appropriate one (for doing so, a user needs to have advanced knowledge on the language). Any comments that might help to understand differences between given language units are very rare.

2. It is very uncommon for dictionaries to present other than lexical ways of expressing certain meaning. As a consequence, if a language B does not have any lexical means to express a certain meaning from language A, the language unit of A bringing this meaning is translated by language units of B that is not a real counterpart.

3. As pointed out in e) above, it is not rare in lexicography to present a language unit A as equivalent of language units B and C without stating expressis verbis if i) the unit A has such a broad meaning that it covers both meaning of B and meaning of C or ii) there are two language units of A's form having two different meanings, out of which one corresponds with B's meaning, second – with C's meaning.

For a dictionary user as well as for a foreigner trying to learn another language it is very difficult to successfully approach the problems mentioned here.

Therefore a parallel corpus providing contexts of usage in two and more languages is a valuable tool. Some examples will be delivered in the process of presenting the problems.

References

Bernardini, S. 2011. "Monolingual comparable corpora and parallel corpora in the search for features of translated language". *SYNAPS – A Journal of Professional Communication* 26: 2-13.

Bogusławski, A. 1995. "Bilingual general purpose dictionary. A draft instruction with commentaries". In J. Wawrzyńczyk (ed.) *Bilingual Lexicography in Poland. Theory and Practice*. Warszawa: Katedra Lingwistyki Stosowanej Uniwersytetu Warszawskiego: 15-55.

Bralewski, D. 2012. *Od przekładu do słownika. Korpus równoległy w redakcji słowników tłumaczeniowych*. Łask: Oficyna Wydawnicza Leksem.

Garabík, R., Dimitrova, L. and Koseska-Toszewa, V. 2011. "Web presentation of bilingual corpora: Slovak-Bulgarian and Bulgarian-Polish". *Cognitive Studies = Études cognitives* 11: 227-239.

Grochowski, M., Kisiel, A. and Żabowska, M. 2010. „Über die Grundsätze der Beschreibung von Stichwörtern in einem zu konzipierenden Wörterbuch der polnischen Partikeln". In L. Zieliński, K. D. Ludwig and R. Lipczuk (eds.) *Deutsche und polnische Lexikographie nach 1945 im Spannungsfeld der Kulturgeschichte*. Frankfurt a.M.: Peter Lang Verlag: 115-130.

Lewandowska-Tomaszczyk, B. 2008. *Corpus Linguistics, Computer Tools, and Applications - State of the Art: Palc 2007*. Frankfurt am Main: Peter Lang Gmbh.

Gains and pitfalls of sentence-splitting in English-Russian translation

Maria Kunilovskaya
Tyumen State University
mkunilovskaya@gmail.com

Natalia Morgoun
Moscow State University
morgounn@yahoo.com

1 Motivation, research data and tasks

In our experience of translator training, one of the major problems is lack of textual cohesion in translations. Most mistakes in text structure are down to the tendency for students to ignore textual features of the source and to translate at best at sentence level. As a result the target text lacks texuality or texture defined by Halliday and Hasan with reference to relations that obtain across sentence boundaries (Halliday and Hasan 1976).

This article focuses on a one of the sources of "cohesion mistakes" in translation, namely those that are associated with sentence-splitting in translation. For the purposes of the present study we define it as change of sentence boundaries, i.e. rendering of one sentence with two or more. By a sentence here we mean a formal graphical sentence running from a capital letter to a full stop and set off by spaces.

The research is based on the data from Russian Learner Translator Corpus (http://www.rus-ltc.org/), It is an on-line parallel corpus of student translations. Its English-Russian subcorpus contains over 200 English non-fiction source texts and their respective multiple 900 translations. The statistics for splitting sentences in the Corpus informs that this transformation is employed in translation of about 5 per cent of source sentence-segments.

Based on the semantic and pragmatic contextual analysis of over 400 English sentences that were split in their Russian translations, this paper aims to describe types of syntactic structures that undergo splitting, along with their semantic and pragmatic properties, typical motivations and results of this shift in English-to-Russian translation. It also contains an overview of typical semantic and pragmatic pitfalls of this shift and attempts to define conditions under which sentence-splitting is justified, as opposed to those, when it is potentially threatening to text cohesion and coherence.

2 Why sentences get split in translation

Detailed analysis of sentences which undergo splitting shows that this technique is almost equally often employed to do away with structural complexity arising from coordination and

subordination.

It turns out that among coordinated structures splitting is most often resorted to when it comes to translating sentences with asyndeton, formally marked by either a semi-colon or a comma, and interclausal "and". Apart from semantic and frequency differences between English and Russian coordinators, we have found out that their range and scope, as sequential discourse markers, differ in that the Russian language more often relies on juxtaposition of sentences for topic continuation. Most of the mistakes here arise from misinterpretation of type of sequential relations signalled by original markers or the scope of their operation.

When it comes to sentences with subordination, splitting results in upgrading of a clause, phrase or verbal or nominative construction to a separate sentence. We offer a frequency order of such structures which is headed non-defining relative clauses and participial and absolute nominative constructions, non-existent in Russian.

It seem that the Russian language does not favour jamming relatively independent additional information into the sentence structure, and therefore, this type of splitting can be typologically justified, especially if the information from the relative clause is continued in the text below. Splitting can also be used to signal discourse relations between bits of information explicitly, which results in a better structured text.

On the whole our statistics shows that in 65 per cent of cases from our data sentence splitting has done no harm to overall translation quality.

3 Typical cohesion and coherence mistakes arising from sentence splitting

Following the Segmented Discourse Representation Theory (Asher and Vieu 2005; Vieu, 2009), we have analysed semantic and discourse relations of the source and target segments in question and arrived at the conclusion that splitting can be potentially dangerous on three counts. It can be effected with disregard to semantic relations between propositions or misinterpretation of the former, including erroneous rendering of semantic connections between proposition by the means of the pragmatic level, for one. Secondly, as this shift requires introduction of a separate sentence, there are problems with its theme and rheme structure. The discourse structure damage to the target is also associated with anaphor resolution which can arise from careless splitting. And finally, there is the effect of a greater communicative value acquired by upgraded sentences which harms the natural flow of information in the text. It is especially dangerous when the information from the element-to-be-a-sentence is not taken on in the subsequent discourse.

References

Asher, N. and Vieu, L. 2005. "Subordinating and coordinating discourse relations". *Lingua* 115: 591-610.

Halliday, M.A.K. and Hasan R. 1976. *Cohesion in English*, London: Longman Group.

Vieu, L. 2009. *Representing Content Semantics, Ontology, and their Interplay*. PhD thesis, Institut de Recherche en Informatique de Toulouse. Available online at http://www.irit.fr/publis/LILAC/LV-HDR09.pdf

Legislative register analysis of Croatian and Italian: intralingual, interlingual and translational perspectives

Ivana Lalli Paćelat
University of Pula
ilalli@unipu.hr

Marko Tadić
University of Zagreb
marko.tadic@ffzg.hr

1 Introduction

Translation and contrastive linguistic studies have significantly benefited from corpora and multilingual corpora in particular (McEnery and Xiao 2008: 18).

It is probably not very well known that in 1968 the usage of computer parallel corpus in contrastive research in the entire history of linguistics was pioneered by Rudolf Filipović in Croatia (Tadić et al. 2012: 76). Although the first English-Croatian parallel corpus was compiled only a year after the publication of the Brown corpus (Kučera and Francis 1967), large parallel corpora for Croatian are still missing (Tadić et al. 2012: 77). Building a large Italian-Croatian parallel corpus of EU legislation has been enabled by the availability of the Croatian translations of the Acquis Communautaire and the possibility to align it further with the JRC-Acquis (Steinberger et al. 2006).

2 Corpus-based translation and contrastive studies

Corpus based translation studies has shown that a translated text differs from a non-translated text and that, independently of the language, translations share some properties (e. g. Baker 1996; Bernardini 2011; Laviosa 2002; Xiao 2010). Whether absolute universals exist or just general tendencies in translated texts is still largely debated (cf. Bernardini and Zanettin 2004; Chesterman 2004; Mauranen 2008; Teich 2003; Xiao 2010; Xiao and Dai 2014).

Research has also been conducted on the differences between registers in translated and non-translated texts and across languages proposing different methodological approaches (e.g. Biber 1995; Neumann 2010; Teich 2003). Biber (1995: 363) holds out 'the possibility of patterns of register variation across languages'. The legal register is on the one hand defined as one of the most 'national registers' (Cortelazzo 1997: 37) which is 'culture dependent' (Engberg 2006: 68), and on the other hand it tends to display universal character, known as 'legalese' (cf. Novak 2010: 3; Tiersma 2006: 552).

3 Aim of the research

The aim of the research is to depict lexico-grammatical features of legislative registers of Croatian and Italian and to compare them in order to find similarities and see whether legislative registers have indeed some universal features. Furthermore, the research aims at finding out whether the translations have the same lexico-grammatical features as the target language legislative register or they belong to a special register. The hypothesis predicts that, given the nature of the legislative register, the lexico-grammatical features are the similar in both languages, no matter how high the frequency of feature occurrence in the reference corpora are. Given the existence of universal translation features, it is assumed that the translated texts are more similar to one another than parallel texts of related languages.

4 Methodology and corpus design

The basic requirements for the register analysis according to Biber (1995) are the comparative approach, the quantitative analysis and a representative sample.

In order for these requirements to be met, six corpora belonging to four different corpus types are employed for the study; firstly, reference corpora for both languages: (1) Croatian National Corpus (HNK v 3.0) and (2) Corpus di Italiano Scritto (CORIS); secondly, (3) specialized bilingual comparable corpus composed of national legislative documents in both languages (subcorpora of HNK v3.0 and CORIS); thirdly, (4,5) monolingual corpora of original national legislative documents and translations of legislative documents of the European Union in the same language used as comparable corpus and lastly, a (6) parallel corpus consisting of Croatian and Italian translations of legal documents of the European Union. For the description of corpus parameters for HNK see Tadić (2002, 2009) and for CORIS Rossini Favretti et al. (2002).

The approach adopted in this study is a hybrid one, without an 'a priori' established theoretical framework, but the corpora are annotated at part of speech (PoS) and lemma level. The analysis is performed by using WordSmith tools v_6.0 (Scott 2013), NoSketch Engine (Rychlý 2007) for HNK v3.0 (Tadić 2009) and for CORIS the on-line interface designed by F. Tamburini. Linguistic feature selection for the quantitative analysis follows previous studies (e.g. Biber and Conrad 2009; Cortelazzo 2013; Rovere 2005; Teich 2003; Venturi 2011; Xiao and Dai 2014), and is driven by primary

corpus obtained data. In order to investigate the properties of translated texts, considered as a special register type, and to find out if there exist universal features of legislative texts across different languages, linguistic features at both lexical and grammatical level are quantitatively analysed and statistically evaluated among all the corpora and the two languages in question.

5 Conclusion

The results showed that the legislative registers of Italian and Croatian share some universal features known as 'legalese'.

While greater similiraties were found, for example, in the distribution of parts of speech, less correspondence was noticed in grammatical means for expressing impersonality and nominal style. Hence, the results of this study confirm that the two languages share the same features of the legislative register, which need not necessarily be expressed by the same grammatical means. However, even at this level, the correspondence was noticed in the majority of cases.

Translational corpora in both languages show the existence of universal translation features, but not always the same features and not with the same frequency (the Italian translational corpus shows the tendency towards normalization and the Croatian translational corpus towards levelling out). However, these features do not make the translations considerably different from comparable original texts in the same language. The results show the largest number of similarities between specialized and translational corpora in the same language, which confirms the authenticity of the translations and their orientation towards the target language, and in particular, towards the features of the target register.

References

Baker, M. 1996. "Corpus-based Translation Studies: The challenges that lie ahead". In Somers, H. (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, (175-187). Amsterdam: John Benjamins.

Bernardini, S. 2011. "Monolingual comparable corpora and parallel corpora in the search for features of translated language", *SYNAPS*, 26, 2–13.

Bernardini S. and Zanettin F. 2004. "When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals". In Mauranen, A., and Kuyamaki, P. (eds.), *Translation Universals: Do they Exist?,* (51-62). Amsterdam: John Benjamins.

Biber, D. 1995. *Dimensions of register variation: a cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, D. 2009. "A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing". *International Journal of Corpus Linguistics,* 14 (3), 275-311.

Chesterman, A. 2004. "Beyond the particular". In Mauranen, A. and Kujamäki, P. (eds.), *Translation Universals: Do They Exist?* (33-49). Amsterdam: John Benjamins.

Cortelazzo, M. A. 1997. "Lingua e diritto in Italia. Il punto di vista dei linguisti". In Schena L. (ed.), *La lingua del diritto: difficoltà traduttive e applicazioni didattiche*, (35-50). Milano: Università Bocconi, Centro linguistico.

Cortelazzo, M. A. 2013. "Leggi italiane e direttive europee a confronto". In *Realizzazioni testuali ibride in contesto europeo. Lingue dell'UE e lingue nazionali a confronto*, Trieste: EUT - Edizioni Università di Trieste, 57-66.

Engberg, J. 2006. "Languages for Specific Purposes". In Brown, K. (ed.), *Encyclopedia of Language and Linguistics 2*. Ugd, (679-683). Oxford: Pergamon Press.

Kučera, H. and Francis, W. N. 1967. *Computational Analysis of Present Day American English.* Providence, RI: Brown University Press.

Laviosa, S. 2002. *Corpus-based Translation Studies: Theory, Findings, Applications.* Amsterdam/Atlanta: Rodopi.

Mauranen, A. 2008. "Universal tendencies in translation". In Anderman, G. and Rogers, M. (eds.), *Incorporating Corpora. The Linguist and the Translator*, (32-48). Clevedon: Multilingual Matters.

McEnery, T. and Xiao, R. 2008. "Parallel and comparable corpora: what is happening?". In G. Anderman and M. Rogers (eds.) *Incorporating Corpora: Translation and the Linguist*, (18–31). Clevedon: Multilingual Matters.

Neumann, S. 2010. "Quantitative Register Analysis Across Languages". In Swain, E. (ed.), *Thresholds and Potentialities of Systemic Functional Linguistics: Multilingual, Multimodal and Other Specialised Discourses*, (85-113). Trieste: EUT Edizioni Università di Trieste.

Novak, B. 2010. *Funkcionalna stilistika hrvatskoga zakonodavstva*. Unpublished PhD thesis, Zagreb: Faculty of Humanities and Social Sciences, University of Zagreb.

Rossini Favretti, R. Tamburini, F. and De Santis, C. 2002. "CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model". In Wilson, A., Rayson, P., and McEnery, T. (ed.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, (27-38). Munich: Lincom-Europa.

Rovere, G. 2005. *Capitoli di linguistica giuridica: ricerche su corpora elettronici*. Alessandria: Edizioni dell'Orso.

Rychlý, P. 2007. "A Modular Corpus Manager". In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, (65-70). Brno: Masaryk University.

Scott, M. 2013. *WordSmith Tools Manual, version 6*, Liverpool: Lexical Analysis Software.

Steinberger R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. and Varga, D. 2006. "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages". In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC2006). Genoa, Italy, 2142-2147.

Tadić, M. 2002. "Building the Croatian National Corpus." *LREC2002 Proceedings*, Las Palmas-Pariz, Vol. II, 441-446.

Tadić, M. 2009. "New version of the Croatian National Corpus". In Hlaváčková, D., Horák, A., Osolsobě, K., and Rychlý, P. (eds.), *After Half a Century of Slavonic Natural Language Processing*, (199-205). Brno: Masaryk University.

Tadić, M., Brozović-Rončević, D. and Kapetanović, A. 2012. *Hrvatski jezik u digitalnom dobu.-The Croatian Language in the Digital Age.* Heidelberg: Springer.

Teich, E. 2003. *Cross-linguistic variation in system and text.* Berlin & New York: Mouton de Gruyter.

Tiersma, P. 2006. "Languages for Specific Purposes". In Brown, K. (ed.), *Encyclopedia of Language and Linguistics, 2. udg*., (679-683). Oxford: Pergamon Press.

Venturi, G. 2011. *Lingua e diritto: una prospettiva linguistico-computazionale*. Unpublished PhD thesis, University of Turin. Available online at: http://www.italianlp.it/?page_id=81 [15.09. 2013.].

Xiao, R., and Dai, G. 2014. "Lexical and grammatical properties of Translational Chinese: translation universal hypotheses reevaluated from the Chinese perspective". *Corpus linguistics and linguistic theory*.

Xiao, R. 2010. "How different is translated Chinese from native Chinese". *International Journal of Corpus Linguistics*, 15(1). 5–35

A corpus-based study of the translation of Chinese "Bei" and "Ba" constructions: Insights from a balanced parallel corpus

Dechao Li
Hong Kong Polytechnic University
ctdechao@polyu.edu.hk

Translating is a kind of mediated communication. As a result, the effect of the source language on the translation is strong enough to make the translational language perceptibly different from the target native language. Translational language can at best be viewed as an unrepresentative special variant of the target language (McEnery & Xiao 2007). The degree of deviation of the translational language can be assessed by studying the distinctive features of the translational language on the basis of contrastive analyses of translated texts and their comparable native texts in the target language (i.e. using the comparable corpus approach), while the extent of source language "shining through" in translations can be identified by comparing the source texts and their translations (i.e. using the parallel corpus approach).

The present study aims to explore the distribution, grammatical features of the translational "Bei" and "Ba" constructions in a balanced parallel corpus from English to Chinese. It is widely believed that "Bei" constructions are "the most typical and frequently-used markers for passive voice in Chinese" (Xiao 2012: 114), which is usually used to describe a state of "unhappiness" or "unwillingness". However, a preliminary study based on the parallel corpus indicates that a majority of the translational "Bei" constructions do not carry with them these associations.

Same as "Bei" constructions, "Ba" constructions are also typical and frequently-used Chinese structures. But unlike "Bei" constructions, they can hardly find any equivalent ones in English. According to Ke (2003: 1), "Ba" constructions are usually used to "to move up an object that can't be placed at the end of the sentence", "to highlight an object so as to emphasize the act and consequences to which the object is related" and "to facilitate the cohesion of the sentences". The study coincides with Ke's study in the first and second aspects, but also finds more functions of translational "Ba" constructions which are rarely seen in spontaneously-written Chinese articles.

The study also attempts to discover to what extent these constructions are influenced by the source language, namely, English, by looking at the general

patterns in which these constructions are produced.

The corpus to be used in the study is a 1 million words English-to-Chinese balanced parallel corpus which covers the same genres as included in the FLOB corpus of British English (Hundt et al 1998). It comprises five hundred 2,000-word text samples proportionally taken from fifteen written text categories (see Table 1). The sampling period of the corpus was from 1991 to 2001.

Register	Code	Text category	No. of samples	Proportion
News	A	News reportage	44	8.8%
	B	News editorial	27	5.4%
	C	News review	17	3.4%
General prose	D	Religious writing	17	3.4%
	E	Skills, trades and hobbies	38	7.6%
	F	Popular lore	44	8.8%
	G	Biography and essays	77	15.4%
	H	Reports and official documents	30	6%
Academic	J	Science (academic prose)	80	16%
Fiction	K	General fiction	29	5.8%
	L	Mystery and detective fiction	24	4.8%
	M	Science fiction	6	1.2%
	N	Adventure fiction	29	5.8%
	P	Romantic fiction	29	5.8%
	R	Humour	9	1.8%
Total			500	100%

Table 1. Corpus design

References

Hundt, M., Sand, A. & Siemund, R. 1998. Manual of Information to Accompany the Freiburg-LOB Corpus of British English. Freiburg: University of Freiburg.

Ke, F. 2003. "The features, distribution and translation of "Ba" constructions in Chinese". *Foreign Language Teaching and Research* 12: 1-5.

McEnery, T. & Xiao, R. 2007. Parallel and comparable corpora: What is happening? In M. Rogers and G. Anderman (eds) *Incorporating Corpora. The Linguist and the Translator*. Clevedon: Multilingual Matters, 18-31.

Xiao, Z. 2012. Corpus-based Studies of Translational Chinese in English-Chinese Translation. Shanghai: Shanghai Jiaotong University Press.

Explicitation and translator's voice: A corpus-based study of multi-item strings in the two English translations of *Hongloumeng*

Liu Kanglong
Hong Kong Shue Yan University
liukanglong@gmail.com

Acclaimed as one of the Four Great Classical Novels of Chinese literature, *Hongloumeng* by Cao Xueqin has long taken a special place among literary scholars and researchers. For a number of decades, translation researchers have been keen on investigating its two full-length translations, one by Hawkes and Minford (*the Story of the Stone*, *Penguin*, 1973-1986) and the other by Yang Xianyi and Gladys Yang (*A Dream of Red Mansions1*, Foreign Languages Press in Beijing, 1978-1980) and have resulted in a number of insightful findings through comparative studies of both works. (cf. Feng 2006; Li et. al. 2011; Wang 2001) The proposed study is to investigate systematically the formulaic languages (multi-item strings) of both translations using a corpus-based approach. As argued by Tannen (1989:37), "[P]re-patterning (or idiomaticity, or formulaicity) is a resource for creativity. It is the play between fixity and novelty that makes possible the creation of meaning". In this study, the 3-word and 4-word multi-item strings are extracted and systematically analysed to shed light on how the two translations differ in its use of translation str3333 ategies. The tentative findings show that Hawkes' translation tends to be more coherent and explicit as it uses many more multi-items (time discourse markers in particular) than the one by the Yangs. Issues regarding repetition, formulaicity and explicitation will also be addressed in relation to the statistical evidence.

The structural and semantic properties of light verb constructions in translational Chinese: A comparison between spoken and written text types

Lu Lu
School of Oriental and African Studies
lu_lu@soas.ac.uk

Xin Huang
Beijing University of Chemical Technology
huangcheers@163.com

The term "light verb", first introduced by Jespersen (1965), refers to the verbs found in expressions whose action is actually described by the nominal object, such as *have a bath*, *take a drive*, and *give a push*. The defining characteristic of these expressions is that the semantic content of the predicate is provided not by the verb, but by its complement. For example, *John gave Lucy a kiss* roughly means *John kissed Lucy*. In examples like this, *give*, the light verb, does not have independent semantic content, which means that any thematic role such a verb has must be semantically vacuous.

Since Jespersen's (1965) coinage, light verb constructions have attracted much attention. In Mandarin Chinese (Chinese, for short), Yin (1980) and Zhu (1982) are acknowledged to be the first researchers to address this issue; they include such words as 进行 *jinxing* 'do' and 加以 *jiayi* 'give'. Though much work has been carried out to look into the interface between syntax and semantics, very limited studies investigate the structural and semantic patterns of light verb constructions in translational Chinese from different text types, i.e. spoken and written translational Chinese. This study thus attempts to demonstrate the contrastive features of Chinese light verb construction in spoken and written Chinese which is translated from English, in order to reveal the structural and semantic properties of Chinese light verb constructions in translated texts and their influences on Baker's (1993) 'universal features of translation'.

In light of the research goal, the sentence-aligned corpora used in this study are Beijing Foreign Studies University Chinese/English Parallel Corpus (CEPC) (Wang 2004) of 5 million characters/words and Corpus of TED Speeches [14] of 6.2 million characters/words for written and spoken translational Chinese, respectively. This paper, according to Xu and Lu's (2013) classification and selection of Chinese light verbs, further explored the structural and semantic properties of 进行 *jinxing* 'do' and 搞 *gao* 'do' of Do group and 加以 *jiayi* 'give' and 给予 *jiyu* 'give' of Give group in two different text types. The translational (non-)correspondences of the Chinese light verb constructions were addressed from two perspectives: a) the syntactic structure and argument structure of light verb constructions and b) the contrastive distribution in spoken and written translational Chinese. All the structural makeup and semantic pattern of light verb constructions between Chinese and their English translations were manually annotated and thoroughly checked.

Results of normalised frequency of translation correspondence demonstrate that Chinese light verb constructions are dominantly (76%) translated from English verbal structures, such as the verbal complement in light verb constructions and light verb constructions themselves, in spoken texts. For example, 进行维修 *jinxing weixiu* 'do repair' prefers to be translated from English verbal construction: 'repair' or 'do the repair' in spoken texts, while it is not the case in written texts. The normalised frequency in different text types show that 进行 *jinxing* 'do' and 给予 *jiyu* 'give' occur more frequently in spoken texts than in the written one in translational Chinese, which contradicts the view proposed by many other researchers (see Diao 2004, for details) that these two words, 给予 *jiyu* 'give' in particular, are a prominent feature of written Chinese.

Apart from the overall translational features across different text types, spoken and written translational Chinese exhibit different syntactic and semantic non-correspondences, especially in passivisation and the addition or omission of certain lexical items. Chinese light verb constructions are more likely to be translated from the passive forms in spoken English than in written one (28% vs. 19%), which is in line with the preferred uses in spoken Chinese, the target language. However, in spoken texts, when translated into Chinese, the omission of subject and object seems to be the prominent feature. For example, in 'beekeepers can replace them very quickly', the translator is inclined to omit the object *them*[15]. The overwhelming feature in translational Chinese is not in accordance with Baker's (1993) universal features of explicitation in translation. Like the syntactic features, the semantic non-correspondence which is addressed from argument structure suggests that semantic explicitation is not a universal feature, especially in the cases of 加以 *jiayi* 'give', 搞 *gao* 'do' and 进行 *jinxing* 'do'. Such sentences as 'if we continue through the entire stack'

[14] Xu, J. 2012. Corpus of TED Speeches. Beijing Foreign Studies University. Available online at http://124.193.83.252/cqp/.

[15] The Chinese correspondence of the sentence in question is '养蜂人当然能很快进行补充.'

is likely to be translated into Chinese in which the agent 'we' is omitted[16]. This study further put forward possible justifications for the features observed.

In summary, this parallel corpus-based study explores the properties in translational Chinese from the analysis of structural and semantic features of light verb constructions across spoken and written text types. This study is significant in exploring the contrastive and translational features of light verb constructions in spoken and written texts. More importantly, the findings give new insight into the properties of light verb constructions and the universal features of translation.

References

Baker, M. 1993. "Corpus linguistics and translation studies: implications and applications." In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and technology: in Honour of John Sinclair*. Amsterdam & Philadelphia: John Benjamins.

Diao, Y. 2004. *Xiandai hanyu xuyi dongci yanjiu* [The study of weak verbs in modern Chinese]. Dalian: Liaoning Normal University Press.

Jespersen, O. 1965. *A modern English grammar on historical principles (volume VI: morphology)*. London: George Allen and Unwin.

Yin, S. 1980. "*Tan jinxing lei dongci weiyuju"* [On the group of predicates like *jinxing*]. In S. Yin (ed.) *Hanyu yufa xiuci lunji* [The selected papers of grammatical rhetoric of Chinese]. Beijing: China Social Sciences Press.

Wang, K. 2004. *Shuangyu duiying yuliaoku yanzhi yu yingyong* [The creation and application of bilingual parallel corpus]. Beijing: Foreign Language Teaching and Research Press.

Zhu, D. 1982. *Yufa jiangyi* [The lectures on grammar]. Beijing: The Commercial Press.

Xu, J. and Lu, L. 2013. "The structural and semantic analysis of the English translation of Chinese light verb constructions: a parallel corpus-based study." In A. Hardie and R. Love (eds.) *Proceedings of Corpus Linguistics 2013*. Lancaster: UCREL. Available online at http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf

[16] The Chinese correspondence of the sentence in question is '如果对一整叠切片进行处理.'

Gender differences and pragmatic markers in conference interpreting

Cédric Magnifico
Ghent University
Cedric.Magnifico@UGent.be

Bart Defrancq
Ghent University
Bart.Defrancq@UGent.be

1 Introduction

This paper is part of a broader research project on gender dimensions of simultaneous interpreting. It focuses on possible gender differences in the use of pragmatic markers by professional interpreters. Pragmatic markers are taken in a broad meaning similar to the approach developed in Brinton (1996). The study of pragmatic markers is particularly relevant in this respect. On the one hand, studies on spontaneous speech have repeatedly shown that women use more pragmatic markers in the form of hedges than men (Lakoff 1975, Homes 1990, Coates (1993,1996). One study on court room interpreting even noted that female interpreters tend to add politeness markers, such as *please* to their interpretations, while male interpreters tend to omit pragmatic markers, such as *well* (Mason 2008). On the other hand, simultaneous interpreting as a linguistic activity is subject to powerful norms (Harris 1990), especially with regard to the faithfulness and completeness of the interpretation. As the aim of norms is to regulate behaviour and, especially, to reduce natural variety in behaviour, simultaneous interpreting is the one linguistic activity in which gender differences should play no or little role. It is therefore an ideal linguistic genre to empirically test the resilience of gender aspects of human speech in the face of norms.

Based on what we know about the interpreting process, we can formulate the following hypotheses:

(1) interpretations are expected to contain fewer pragmatic markers than the source text: simultaneous interpretation is an extremely demanding cognitive task and interpreters are trained to give the propositional meaning of the source utterance priority (Seleskovitch 1975). Pragmatic markers are not part of that propositional meaning (Fraser 1999) and will therefore be more often omitted if the interpreters face cognitively demanding source texts;

(2) interpretations carried out by female interpreters are expected to contain more pragmatic markers than interpretations carried out by male interpreters, as women use at least some categories of pragmatic markers more often than men.

2 Data

The EPICG (European Parliament Interpreting Corpus Ghent) corpus has been compiled at Ghent University and is based on plenary sessions held at the European Parliament in 2006 and 2008. The corpus comprises 193,000 words, including source speeches in French, Spanish and Dutch and their interpreted versions in Dutch, English and French. For the research on PMs, we have selected the sub-corpus where French is the source language and Dutch and English the target languages (147,000 words).

The transcriptions also include a large number of oral features, i.e. hesitation markers, false starts, repetitions and so forth. Each speech displays metadata, specifying the name of the speaker, the topic, the date, the duration, the number of words and the interpreter's gender.

3 Methodology

All pragmatic markers in source and target texts were identified manually both in the source texts and in the target texts. Source and target texts were then compared and occurrences of pragmatic markers were classified into two categories, depending on the relation between source and target texts: (1) PMs involved in a translation relationship, i.e. when both the source and target texts contain PMs at similar positions in the utterance; (2) PMs not involved in a translation relationship. The second category was further split into PMs occurring in a source text without an equivalent in the target text and PMs occurring in a target text without an equivalent in the source text. Occurrences were then cross-classified according to interpreter's gender.

4 Results

The first – rather surprising – result of the analysis is that target texts always contain more PMs than source texts. Interpreters tend to add PMs to their interpretations, especially markers of discourse structure such as additive *ook* ('also') and forward causal *dus* ('so'), which have no counterpart in the source text in about 50% of the cases. Pragmatic markers are also omitted, which leads to the interesting conclusion that the use of PMs in source and target texts only overlaps to a very limited extent. With regard to gender differences, female interpreters are found to be less prone than male interpreters to omit pragmatic markers occurring in the source text. They also appear to add more markers, confirming earlier findings on gender-biased marker usage in spontaneous speech. There also seems to be a bias at the level of the individual markers: some markers are predominantly used by women (Dutch *nou*, for instance) and some by men. This also confirms earlier findings by *inter alia* Andersen (2001) on the basis of spontaneous spoken language. However, the markers with a strong gender bias in our study belong to different categories than the ones mentioned in previous research.

References

Andersen, G. 2001. Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents. Amsterdam: John Benjamins.

Brinton, L.J. 1996. *Pragmatic Markers in English, Grammaticalization and Discourse Functions*. Berlin: Mouton de Gruyter.

Coates, J. 1993. *Women, men and language*. 2nd edition. London: Longman.

Coates, J. 1996. "You know so I mean probably: Hedges and hedging". In: Coates, J. (ed.) *Women Talk: Conversation Between Women Friends.* Oxford: Blackwell. 152–173.

Fraser, B. 1999. "What are discourse markers?". *Journal of Pragmatics* 31: 931-952.

Harris, B. 1990. "Norms in Interpretation", *Target*, 2: 115-19.

Holmes, J. 1990. Hedges and boosters in women's and men's speech, *Language & Communication*, 10(3): 185-205.

Lakoff, R. 1975. *Language and Women's Place.* New York: Harper Colophon.

Mason, M. 2008. *Courtroom Interpreting*. Lanham: University Press of America.

Moser-Mercer, B. 1978. "Simultaneous Interpretation: a Hypothetical Model and its Practical Application". In D. Gerver et H.W. Sinaiko (eds.) *Language Interpretation and Communication.* New York/London: Plenum Press.

Seleskovitch, D. 1975. *Langage, langues et mémoire, étude de la prise de notes en interprétation consécutive.* Paris : Minard Lettres Modernes.

"NP internal" *kind of*: Evidence from a parallel translation corpus

Michaela Martinková
Palacky University
michaela.martinkova@upol.cz

Markéta Janebová
Palacky University
marketa.janebova@upol.cz

1 Introduction

Recently, a lot of attention has been paid to the patterns with nouns originally denoting type or subclass, namely *sort*, *kind* and *type.* Aijmer (2002, 176) differentiates between the pattern (exemplified in 1) in the hyponymy statement *robin is a sort* (N1) *of bird* (N2), where the noun *sort* is the head of the NP and the *of*-phrase its modifier (Denison 2005 talks about a "binominal construction", Davidse et al. 2008 about a "lexical head use"), and (2), where "*sort of* modifies the nominal head":

(1) *can you just tell me what sort of unit trusts they are what sort of industries they're invested in*

(2) *there are these sort of practical problems*

The incongruence in number between *sort* and the determiner in (2) is taken as formal evidence of the modifier status of *sort of*. The fact that spoken language data are subject to scrutiny (LLC) allows Aijmer to study not only the grammatical behaviour of *sort of/kind of* and their collocational patterns, but also their prosodic features: e.g., *sort of/kind of* in the modifier use is always unstressed. *Sort of* in (3) is then given as an example of its use as a discourse particle – it is followed by a pause and has a metalinguistic (hedging) function (182):

(3) *one can imagine a sort of middle-age woman*

Denison (2005) suggests a hedging function also for (4), which he posits as an instance of a "qualifying construction" (with N2 as the head):

(4) *When thanks is not forthcoming, we feel a kind of emptiness*

Davidse et al. (2008) further elaborate on Denison's taxonomy. For example, they extend the "postdeterminer" or "complex-determiner construction", which Denison posits only tentatively, to cover other cases (often analogous to *such* in its anaphoric function).

This pilot study focuses on NP internal uses of the type nouns and their functions, but uses a different methodology. Following Johansson, we turn to a parallel translation corpus to investigate those meanings of type nouns which are "visible through translation" (2007, 57). The language selected is a typologically distant language (Czech).

2 Data and methods

The data come from Intercorp, a multilingual translation corpus of Czech and 31 languages. The present structure of the corpus forces us to focus only on American English: a subcorpus of post-1920 American fiction (3,278,423 words) and its Czech translations was created and all tokens of *kind of* immediately followed by a noun downloaded (521 tokens). *Kind of*, favoured by American English (e.g. Denison 2005 and Biber et al. 1999, 871), was then subjected to a deeper analysis. In all but one token it was indeed a part of an NP.

3 Discussion of findings

A type noun (such as *druh* "kind" and *typ* "type") was found in the translation of 88 tokens, i.e. 16.9% of all the tokens of *kind of*. Syntactic restructuring (a verb is used in Czech) also allows expressing the type noun as a different POS (*žánrově* "as far as genre is concerned").

However, even Czech type nouns can function as hedges, if part of lexicalized phrases (*svého druhu* "of its kind"). Syntactic restructuring allows for the use of the hedging phrase *svým způsobem* ("in its own manner"). The hedging function is also made explicit by the indefinite pronoun *jakýsi*, the imperative phrase *řekněme* ("let's call it"), and a downtoner[17] (*a kind of hiccup – takové téměř škytání* "such almost hiccup"). All of these translations with a Czech hedge, arguably, mark what Denison calls a "qualifying construction" and Davidse et al. (2008) a "nominal qualifier use".

The expression *takový* ("such") is found in 65 tokens (12.5% of all tokens of *kind of*), which confirms parallels suggested in the linguistic literature ("postdeterminer use"). This use may be purely anaphoric (in which case the suffix *hle* (originally "look!") grammaticalized in strong forms of Czech demonstrative pronouns is in two cases added to form the informal *takovýhle*), but in some cases, as Davidse et al. (2008) argue, for pragmatic reasons it may suggest "size intensification", which in turn can get an emotional colouring (*this kind of money – takové peníze* "such money").

A negative emotional colouring can be found in *co je to za* (21 tokens), the equivalent of *what kind of*. However, in 50% of all its tokens *what kind of* is translated just with *jaký* ("what"), which, unlike its English equivalent, covers both the general and specific use of N2. This opens the question of "zero correspondences", i.e. tokens in which no translation equivalent could be identified within the scope of one sentence. In our sample, it covers not only *kind of* preceded by determiners with direct Czech

[17] Downtoners (Quirk et al. 1985) are found in four more cases.

equivalents, but also 30% of all tokens of *the kind of*, in which *the* is cataphoric ("postdeterminer use"), and N2 is followed by a relative clause.

4 Conclusions and looking ahead

Czech translation equivalents show that the type noun *kind* as part of an NP has predominantly a pragmatic function, in which it loses its head status. Apart from this "nominal qualifier use" translations can also make explicit its much less frequent "lexical head use". The "postdeterminer use" has less straightforward equivalents, especially if it has a cataphoric reference. The analysis reveals differences between English and Czech determiners and calls for more research on a monolingual corpus of Czech.

References

Aijmer, Karin. 2002. *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins.

Biber, D., et al. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Czech National Corpus – InterCorp. Institute of the Czech National Corpus, Prague. Available online at http://www.korpus.cz.

Davidse K., Brems, L. and De Smedt, L. 2008. "Type noun uses in the English NP: A case of right to left layering". *International Journal of Corpus Linguistics* 13 (2): 139–167.

Denison, David. 2005. "The grammaticalisations of *sort of*, *kind of* and *type of* in English." A presentation at New Reflections on Grammaticalization 3, Santiago de Compostela. Available online at http://www.humanities.manchester.ac.uk/medialibrary/llc/files/david-denison/Santiago_NRG3_paper.pdf

Denison, David. 2007. "Playing tags with category boundaries." *Varieng: Studies in Variations, Contact and Change in English.* Available online at http://www.helsinki.fi/varieng/series/volumes/01/denison/

Halliday, M. A. K. 1994. *An Introduction to Functional Grammar*. London: Arnold.

Johansson, S. 2007. "Seeing through Multilingual Corpora". In R. Facchinetti (ed.) *Corpus Linguistics 25 Years On*. Amsterdam – New York: Rodopi.

Quirk, R., et al. 1985. *A Comprehensive Grammar of the English Language*. London: Logman.

Comparing literary translations with principal component analysis: A methodological application, its advantages, and its limitations

Lorenzo Mastropierro
University of Nottingham
lorenzo.mastropierro
@nottingham.ac.uk

1 Introduction

Principal component analysis is a multivariate analysis that provides a measure of the overall degree of difference between sets of data, for example whole texts, based on the frequency patterns of a pool of variables, i.e. words. John Burrows is generally regarded as the scholar who introduced multivariate analysis in corpus stylistics and his computational study of Jane Austen's novels (Burrows 1987) has paved the way for further research. Since then, multivariate and principal component analysis have been used extensively in both corpus linguistics and corpus stylistics.

The basic idea behind their application is that by taking into account a wealth of variables – many of which may be weak discriminators –, multivariate analysis provides a more tenable result of the overall texts relation than when a smaller number of stronger discriminators are used (Burrows 2002: 679). However, despite the popularity of this method in stylistics, stylometry and authorship attribution studies, its use in the study of translation – literary translation in particular – has only just begun to emerge (see Rybicki 2006 and Rybicki & Heydel 2013, for example).

2 Aim

This paper aims to show the application of principal component analysis to translation studies, and to discuss its methodological implications. Through a comparative analysis of four Italian translations of Joseph Conrad's *Heart of Darkness* (1899), this paper aims to demonstrate the potential contribution of this procedure to this research context, as well as its limits.

3 Methodology

In the first part, the analysis focuses on comparing each translation with the others, first using single words as variables, then repeating principal component analysis with two-word sequences and three-word sequences. The four translations include, on the one hand, the first Italian translation of *Heart*

of Darkness (1928); on the other, three contemporary translations (1990s). The outcome of these comparisons helps to trace the interrelationships among the translations, revealing the degree of similarity and difference between them, both from a diachronic and a stylistic perspective.

In the second part, the focus is then moved to the relation between the source text and the target texts. In order to do so, both the source text and the target texts are segmented into 10 sections. Principal component analysis is then used to highlight the degree of similarity among the sections. The section-clustering on the resulting score plots serves as shared ground for the comparison of the target texts with the source text.

Finally, the findings of the two parts of the analysis are discussed in relation to each other in order to provide conclusive remarks on the application of principal component analysis to the comparison of literary translations.

4 Expected results

This paper expects to show how principal component analysis can be used to study literary translation. It tries to prove the effectiveness of this statistical procedure in the comparison of texts in a translational context, as well as the limitation of such an application. In light of this, this paper argues for a compensative methodology that links together the multivariate analysis approach with a more bottom-up perspective, such as that provided by a corpus stylistic analysis.

This is argued to result in the intersection between fine-grained examinations, based on individual textual features identified with the help of a corpus stylistic analysis, and a broader perspective on the overall relation between the original and its translation based on their mutual degree of difference, such as that provided by principal component analysis.

References

Burrows, J. 1987. Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Clarendon.

Burrows, J. 2002. "The Englishing of Juvenal: Computational stylistics and translated texts". *Style* 36 (4): 677-750.

McKenna, W., Burrows, J. & Antonia, A. 1999. "Beckett's trilogy: Computational stylistics and the nature of translation". *Revue Informatique et Statistique dans les Sciences Humaines* 35: 151-171.

Rybicki, J. 2006. "Burrowing into translation: Character idiolects in Henryk Sienkiewicz's Trilogy and its two English translations". *Literary and Linguistic Computing* 21 (1): 91-103.

Rybicki, J. & Heydel, M. 2013. "The stylistics and stylometry of collaborative translation: Woolf's *Night and Day* in Polish". *Literary and Linguistic Computing* 28 (4): 708-717.

Genderlect in Enron: a contrastive corpus based investigation of language variance in corporate email

Jamie McKeown
Hong Kong
Polytechnic University
Jamie.mckeown
@gmail.com

Li Lan
Hong Kong
Polytechnic University
Lan.Li
@polyu.edu.hk

1 Introduction and objective

The collapse of Enron in 2002 perhaps secured for the corporation a nonpareil status in its ability to weigh on the global psyche as a symbol of spectacular corporate failure (Swartz and Watkins 2004). The notorious shredding of documents, mark-to-market accounting, the rank and yank employee grading system and the conversion of stock by senior board members ahead of the release of negative results have all been attributed to the work of a few leading males and their encouragement of subordinates (Maclean and Elkind 2004). Despite the fact that the atrophy of the corporation was largely precipitated by the courageous acts of two women: the whistle-blowing senior accountant (Sherron Watkins) and the young maverick journalist Bethany Maclean (whose 2001 Fortune magazine article 'Is Enron overpriced' first dared to question the hubris of the energy behemoth), little attention has been given to the role or even presence of women within the organization (save Playboy Magazine's, post liquidation, feature of 10 female ex-employees).

This study in building on previous work regarding gender and language variation (Tannen 1990; Wodak 1997; Baxter 2003; Koller 2004) will primarily look to explore the degree to which discrete genderlects (Tannen 1990) are evident in the workplace email of the hyper-emasculated context that was Enron. Modern theories of language and gender claim that men dominate interactions with women and the language system itself, whilst the use of language by women carries certain features that mark inferiority (Lakoff 1975). Some studies suggest that men tend to use language instrumentally, while women mainly use language to maintain relationships (Cameron 1995). Men use language in a competitive way, reflecting their supposed interest in acquiring status; women use language in a cooperative way, reflecting their preference for equality and harmony (Holmes 1995). Post-modern theories deny differences in language behaviors as being attributed to gender. Culture, status, and the intent of the communicator, have much more influence on stylistic variations than sex (Mulac 1998; Goddard and Mean 2009). Through the course of this study we will attempt to reveal if there was a separate genderlect in operation in Enron or perhaps if the culture was so pervasive as to furnish little contrast in the communication styles of the respective sexes.

2 Data and approach

The data in the proposed study will be taken from the corpus of 500,000 emails originally made public by the U.S Federal Energy and Regulatory Commission during its investigation of Enron. The emails used in this paper are taken from a subset of 1700 labeled email messages focusing on business-related emails and the California Energy Crises, released by Marti Hearst at UC Berkeley. The data will be split into two sub-corpora: 'En-men' corpus and 'En-women' corpus. In order to protect the privacy of individuals all examples presented will be done so in a redacted form.

Through the use of Wordsmith 5.0 the measure of keyness will be used a method of analysis for the fact that it facilitates the identification of differences between corpora (McEnery and Hardie 2012). As a measure keyness enables the analyst to see which words are used significantly more frequently thus reflecting what the text is truly about (Scott and Tribble 2006) and for our purposes what variations exist between the two sub-corpora. In order to generate the keyness measure one sub-corpus will be used as the reference corpus of the other e.g. En-men will be used as a reference corpus for En-women, thus dispensing with the use of a third reference corpus. In order to detect the similar salient features of the respective corpora the frequency word lists of each data set will be generated and examined for such propensities.

A combination of approaches associated with pragmatics (transitivity, Halliday and Matthiessen 1999; rapport work, Locher and Watts 2005; face and politeness, Brown and Levinson 1987), communication theory (Relational Practice, Holmes 2006) and gender studies (Difference theory, Tannen 1990) will be used when analysing the email texts. When reading the concordance lines of keywords and phraseologies, attention will be paid to the lexico-semantic relations in an attempt to understand the possible motivation and function behind the lexical choices made.

3 Value of the study

The proposed study will hopefully contribute to a number of existing streams of knowledge. Firstly, we hope to add knowledge to an understanding of what took place within the corporate jungle of

Enron. Secondly, we hope to add to the burgeoning field of contrastive language use in workplace communication. Finally the proposed study will be unique in the examination of language and gender in email at the textual level.

References

Baxter, J. 2004. Positioning gender in discourse: a feminist methodology. Palgrave Macmillan.

Brown, P. and Levinson, S.C. 1987. *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Cameron, D. (1995). Rethinking language and gender studies: Feminism into the 1990s. In S. Mills (ed.*), Language and Gender: Interdisciplinary perspectives*. London: Longman.

Cohen, W. 2009. Enron email dataset. http://www.cs.cmu.edu/~enron/, accessed on 30 October 2013.

Goddard, A. and Mean, L. 2009. *Language and Gender*. London: Routledge.

Halliday, M. and Matthiessen, C. 1999. *Construing Experience Through Meaning: A Language-Based Approach to Cognition.* London: Cassell.

Holmes, J. 2006. *Gendered Talk at Work.* Oxford: Blackwell.

Holmes, J. 1995. *Women, Men and Politeness*. London: Longman.

Koller, V. 2004. *Metaphor and gender in business media discourse: a critical cognitive study.* Palgrave Macmillan.

Lakoff, R. 1975. *Language and women's place.* New York, NY: Harper and Row.

Locher, M. and Watts, R. 2005. "Politeness theory and relational work". *Journal of Politeness Research* 1 (1), 9–33.

Maclean, B. 2001. "Enron: Is Enron overpriced?" *Fortune.* 143(5), 123-130.

Maclean, B. and Elkind, P. 2004. *The Smartest Guys In The Room: The Amazing Rise And Scandalous Fall Of Enron.* Portfolio Trade.

McEnery, T. and Hardie, A. 2012. *Corpus Linguistics. Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mulac, A. (1998). The gender-linked language effect: Do language differences really make a difference? In D. Canary & K. Dindia (Eds.), *Sex differences and similarities in communication: Critical essays and empirical investigations of sex and gender in interaction* (pp. 127-153). Mahwah, NJ: Lawrence Erlbaum.

Scott, M. and Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis*. Amsterdam: John Benjamins.

Swartz, M. and Watkins, S. 2004. *Power Failure: The Inside Story of the Collapse of Enron.* Doubleday.

Tannen, D. 1990. *You just don't understand: women and men in conversation*. Morrow.

Wodak. R 1997. *Gender and discourse.* Sage Publications.

Nominalization in literary texts: a corpus-based study of contrastive and translational aspects

Tamara Mikolič Južnič
University of Ljubljana
tamara.mikolic@guest.arnes.si

1 Introduction

The study focuses on nominalization and its occurrence in Italian and Slovene literary texts. The corpus used in the study comprises Italian source texts and their Slovene translations, as well as Slovene original literary texts. Corpus-based research methodology is used to show how the frequency of nominalization seems to be language-dependent, as well as genre-dependent,[18] and how its presence in literary texts seems to be affected through the process of translation. More precisely, we are interested in what way Slovene translated literary texts differ from original ones with regard to the presence of nominalization. Since not all nominalizations occurring in the translated texts are direct translations of Italian nominalizations, also those occurring when the source text uses other means of expression will be taken into account. The aim is to verify what is their overall frequency in the corpus, how often they occur as translations of source text nominalizations and what proportion is the result of other structures in the source texts; finally, we are also interested in what structures are found in the source texts when target text nominalizations are not the result of a direct translation.

From a contrastive point of view, therefore, the structures appearing in Slovene texts in place of the source text nominalizations, and those found in source texts where additional nominalizations are found in the target texts, will be analysed both with regards to their type and their relative frequency. From the viewpoint of translation studies, some possible explanations will be explored concerning the possible reasons behind the difference in frequency of nominalizations in the two languages, i.e. interference (Toury 1995) and explicitation (Klaudy and Karoly 2005).

[18] To show how the occurrence of nominalization in literary texts compares with other genres, the results of the analysis will be compared with previous research (Mikolič Južnič 2007, 2010, 2011, 2012a, 2012b)

2 Nominalization as a grammatical metaphor

In this study, nominalization is viewed in the light of Halliday's systemic functional grammar (Halliday 1994, Halliday and Matthiessen 2004), as a particular type of grammatical metaphor of the ideational plane, whereby a process is realized by a noun.[19] Such realizations cause a rearrangement of the whole sentence structure and the casting of the participants in the underlying process as modifiers in the nominal group. As a result, the sentences are lexically denser, as more information is packed into single units. While this seem perfectly acceptable in Italian, to the average Slovene reader, such a nominally loaded style seems to be difficult to comprehend (cf. Žele 1996) and it is therefore frequently avoided in a number of genres, among which there are literary texts.

3 Corpus and method

The research presented here is part of a wider study on the presence of nominalization in various Slovene and Italian genres (cf. Mikolič Južnič 2007, 2010, 2011, 2012a, 2012b, 2013). It was carried out with the help of the Spook corpus (Vintar 2009), a translation corpus of literary texts that consists of two main sections: original Slovene literary texts and literary texts translated into Slovene from four languages (English, French, German and Italian), as well as the source texts of the translations. As it was mentioned above, only the original Slovene literary texts and the translations from Italian (and their source texts) were used in the study. Slovene nominalizations were identified mostly through relatively simple queries of strings of characters and wild cards; afterwards the concordances were manually checked and analysed in order to determine the relations between the nominalizations and the structures used in the source texts. The results were then compared with those found in original Slovene literary texts and in other written and oral genres.

4 Results

The results show that nominalization is indeed much less present in Slovene literary texts compared to other genres. Its frequency is also lower compared to the occurrences in Italian literary texts. From the viewpoint of the source texts, a number of source nominalizations are not translated directly, therefore a variety of alternative options are given, the most frequent being an explicitation with a finite verb. When observing all the nominalizations present in

[19] In a congruent wording, a process is realized by a verb (cf. Halliday 1994: 343)

the translated Slovene texts, we notice also that there is a considerable number of them resulting from other Italian structures, mostly non-finite verb forms.

References

Halliday, M. A. 1994. *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M. A. and Matthiessen, C. M. 2004. *An Introduction to Functional Grammar. Third Edition.* London: Arnold.

Klaudy, K. and Karoly, K. 2005. "Implicitation in Translation. Empirical Evidence for Operational Asymmetry in Translation". *Across Languages and Cultures* 6 (1): 13-29.

Mikolič Južnič, T. 2007. *Nominalne strukture v italijanščini in slovenščini : pogostnost, tipi, in prevodne ustreznice. Unpublished* PhD thesis, University of Ljubljana.

Mikolič Južnič, T. 2010. Translation of Italian Nominalizations into Slovene: a Corpus-Based Study. *RITT (Rivista Internazionale di Tecnica della Traduzione)* 12: 145-158.

Mikolič Južnič, T. 2011. "Vpliv besedilnih tipov na pojavljanje nominalizacije v slovenščini: korpusna raziskava". In S. Kranjc (ed.), *Meddisciplinarnost v slovenistiki. Obdobja 30.* Ljubljana: Znanstvena založba Filozofske fakultete. 321-327.

Mikolič Južnič, T. 2012a. "A contrastive study of nominalization in the systemic functional framework". *Languages in Contrast 12 (2)*: 251-276.

Mikolič Južnič, T. 2012b. "La nominalizzazione come indicatore del grado di formalità in alcuni tipi testuali della lingua parlata". *Linguistica* 52: 283-295.

Mikolič Južnič, T. 2013. "Bridging a grammar gap with explication : a case study of the nominalized infinitive". *Across Languages and Cultures* 14 (1): 75-98.

Toury, G. 1995. *Descriptive translation studies and beyond.* Amsterdam / Philadelphia: John Benjamins.

Vintar, Š. 2009. "Slovenski prevodoslovni korpus". In M. Stabej (ed.), *Infrastruktura slovenščine in slovenistike*. Ljubljana: Znanstvena založba Filozofske fakultete. 385-391.

Žele, A. 1996. "Razvoj posamostaljenja v slovenskem publicističnem jeziku med 1946 in 1995". In A. Vidovič Muha (ed.), *Jezik in čas*. Ljubljana: Znanstveni inštitut Filozofske fakultete. 191-200.

Expert knowledge representation in general English/Spanish dictionaries: a case study

Maria Teresa Ortego-Antón — tortego@lesp.uva.es

Purificación Fernández-Nistal — purifier@itbyte.uva.es

University of Valladolid, Spain

In recent decades, the number of new concepts and terms has risen rapidly due to scientific and technological development. Additionally, expert knowledge, which used to be exclusive for experts, also interests middlebrow language users as a result of the democratisation of education and the media broadcasting. Compilers of e-dictionaries are aware of this change, so in new editions, they are gathering specialised terms that have become part of our daily lives.

Moreover, in the current globalised world, the need to transfer scientific knowledge to other languages arises, since it is produced or spread mainly in English. In this framework, the transfer of specialised vocabulary is one of the obstacles that translators and, specially, translation trainees deal with, so one of the main tools that they employ to look up an unknown term are bilingual dictionaries. Despite the fact that they are not the most suitable tool to search for specialised vocabulary because they often lead to mistakes when concepts are unknown, in previous research (Atkins&Varantola 1998a, 1998b; Durán Muñoz 2010; Bowker 2012) dictionaries were reported to have become one of the most generalised and frequently used tools among translators and interpreters.

On the other hand, the analysis of the entire specialised vocabulary gathered in bilingual dictionaries is a task previously defined as difficult if not impossible (Thoiron 1998: 624-625; Rodríguez Reina 2002: 352-353). Consequently, our study is limited to a particular field of knowledge, that is, computing. This domain is cross-sectional to other domains, in the sense that nowadays computing applications hold up all the domains in our society, to the point that their changes have an impact on the advances of most of the human areas.

Taking into account the difficulties arising from scientific vocabulary transfer in interlingual communication as well as the importance of bilingual e-dictionaries as a search tool for users, we consider that the study of the treatment given to computing terms in three of the most used bilingual dictionaries (*Collins Universal*, *Gran Diccionario Oxford* and *WordReference*) is a field that needs to be reviewed as long as it can offer resulting data that

might improve the information gathered and implement the search procedures used by translators and interpreters. From an ad hoc corpus composed of texts from the main journals published in the UK and the USA the most frequent terms belonging to computing will be extracted using *TermoStat Web 3.0* (Drouin 2003). This extractor identifies items using a statistical technique that compares frequencies in a technical and non-technical corpus, which are shown in a list From the results offered by *TermoStat Web 3.0*, terms as well as proper nouns are manually revised and excluded. Then, the treatment given to computing terms is analysed following a methodology used in previous research (Roberts 2004; Josselin 2005; OrtegoAntón 2012): first, we verify if the selected terms included in the dictionary wordlist as entries, nest entries or examples, then, if they are labelled with computing, which translation equivalents are given and if they are followed by contextual data. In addition, we will find out if the given equivalents are used in Spanish language checking their use in two Spanish reference corpora: *Corpus del Español* and *Corpus de Referencia del Español Actual.*

The results from the analysis might suggest a need to take into account new proposals in order to implement the data gathered in these reference works as well as inform new procedures in the design and use of these tools from the point of view of translators as main users.

References

Atkins, B. T. S. & K. Varantola. 1998. "Monitoring Dictionary Use". In B. T. S. Atkins (ed.) U*sing Dictionaries: Studies of Dictionary Use by Language Learners and Traslators.* Tübingen: Max Niemeyer Verlag, 83-122.

Bowker, L. 2012. "Meeting the needs of translators in the age of e-lexicography. In S. Granger & M. Paquot (eds.) *Electronic Lexicgraphy*. Oxford: Oxford University Press, 373-391.

Drouin, P. 2003. "Term extraction using non-technical corpora as a point of leverage", *Terminology* 9(1): 99-117.

Durán Muñoz, I. 2010. "Specialised lexicographical resources: a survey of translators' needs". In S. Granger & M. Paquot (eds.) *eLexicography in the 21st century: New Challenges, New applications. Proceedings of ELEX 2009, Cahiers du Centra.* Louvain-la-Neuve: Presses Universtaires de Louvaine-La-Neuve, 55-66.

Josselin, A. 2005. *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues. Étude d'un domaine de spécialité: Volcanologie.* PhD Thesis. Lyon: Université Lumière Lyon II, Centre de Recherche en Terminologie et Traduction. Available online at http://theses.univ-lyon2.fr/documents/lyon2/2005/josselin_a#p=0&a=top

Ortego Antón, M. T. 2012. *Estudio contrastivo inglés/español del tratamiento del léxico especializado recogido en los diccionarios generales bilingües: Collins Universal y Gran Diccionario Oxford.* PhD Thesis. Soria: Universidad de Valladolid. Available online a t https://uvadoc.uva.es/handle/10324/2022

Roberts, R. P. 2004 "Terms in General Dictionaries". In J. M. Bravo Gozalo (ed.) *A New Spectrum of Translation Studies.* Valladolid: Universidad de Valladolid, 121-140.

Rodríguez Reina, Mª Pilar. 2002. "Las marcas de especialidad: Una cuenta pendiente de difícil solución". In Guerrero Ramos & M. F. Pérez Lagos (eds.) *Panorama actual de la terminología*. Granada: Comares, 327-357.

Thoiron, P. 1998. "Place et rôle de la terminologie dans les dictionnaires bilingues non spécialisés. Le cas de la terminologie médicale dans le Dictionnaire Hachette-Oxford (français-anglais)". In S. Mellet and M. Vuillaume (eds.) *Mots chiffrés et déchiffrés*. Paris: Honoré Champion, 621-650.

The challenges of translating specialized collocations and extended collocations in law documents: a corpus-based research

Adriane Orenha-Ottaiano
Universidade Estadual Paulista (UNESP)
adriane@ibilce.unesp.br

1 Introduction

Considering the great relevance of legal and sworn translation in commercial, social and legal relations among nations, a parallel and a study corpus made up of articles of incorporation and bylaws were compiled (as detailed below in the Methodology Section), in order to investigate the use and translation options of specialized collocations in the referred documents. The research also introduces the term 'extended specialized collocations' (Orenha-Ottaiano 2009), that is, specialized collocations which are meant and built in blocks, coined to describe the occurrence of more expanded collocations whose characteristics are inherent to the so-called specialized phraseological units. Both specialized collocations and extended specialized collocations were chosen to be investigated given their recurrent and conventional nature in law documents, besides the difficulty they pose to translators. We strongly believe collocational awareness is highly relevant to learner and professional translators and that the results of this investigation may contribute to a deeper reflection of the role of the referred phraseologisms in translation.

2 Methodology

With a view to extract (extended) specialized collocations from articles of incorporation and bylaws, drawing upon the theoretical and methodological framework of Corpus-Based Translation Studies and Corpus Linguistics (Tognini-Bonelli 2001; O'Keeffe and McCarthy 2010), Phraseology (Bertrand and L´Homme 2000; Hausmann 1985; Meunier and Granger 2008; Orenha-Ottaiano 2009) and studies on sworn translation (Aubert 2004, 2005; Mayoral-Asensio 2003), it was compiled: 1) a parallel corpus of 95,618 words, comprised of articles of incorporation and bylaws submitted to the process of sworn translation in the translation directions from English into Portuguese and from Portuguese into English, excerpted from the Books of Sworn Translation Records, made available by five Brazilian sworn translators, duly sworn by the Board of Trade of two Brazilian States; 2) a study corpus of 298,837 words, made up of translated documents of the same nature submitted and not submitted to the process of sworn translation, in the same translation directions; and 3) two comparable corpora of 396,760 words, composed of the referred documents originally written in Portuguese and in English.

3 Data analysis

According to the data analysis result, many types of specialized collocations were raised, for instance, verbal, nominal, adjectival and adverbial collocations and, some of the collocational options investigated were not frequently found in the target language. Regarding the extended specialized collocations, they were found to be recurrent, stable and conventional lexical combinations, some with a high degree of fixedness, made up of some fixed elements as in *shares that a company purchases, redeems or otherwise acquires may be cancelled or held as treasury shares*. Others appeared to have more variable elements – some may accept suppressions or insertions of components and, in some other cases, may allow a change in the order of their elements. It can hence be argued that linguistic data are not enough to proceed to the identification of extended specialized collocations, as pragmatic aspects need to be considered. The analysis showed that, due to the fact there is a correlation between language and culture, and that this aspect may affect the way one combines words, when the correspondent extended specialized collocations in the target language were analyzed, they seemed not to be frequent and recurrent. Besides this cultural aspect, one should also regard the great difference in the focused law systems (Brazilian and North-American), which may also affect the choices and combination of words in the two languages.

4 Conclusion

Considering the analyzed data, it may be stated that culture is manifested in language and vice-versa. Therefore, culture is manifested in collocations, specialized collocations and extended specialized collocations. Cultural knowledge is intrinsically related to lexical competence, that is to say, the choice of collocations, specialized collocations and extended specialized collocations is restricted to certain cultural stereotypes, once some elements in combinations, due to cultural specificities, differ from a language to another. Hence, lack of cultural and phraseological competence may lead to production of non-fluent texts or translations. That implies that translators, for instance, should translate not only words, but chunks or blocks of words, having in mind the lexical patterns of a language and

its cultural aspects. Based on the types of collocations extracted and the collocational errors detected, it is argued that the translation of specialized collocations, and mainly of extended specialized collocations, may be considered a challenge to both sworn and legal translations, and studies like the one here proposed is believed to be a step towards helping learner and not so experienced translators be aware of them and produce more natural texts.

Acknowledgement

I gratefully acknowledge the financial support provided by *CAPES* (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) that made my participation in this conference possible.

References

Aubert, F. H. 2003/2004. "Dúvidas e controvérsias. Tradução juramentada: qual a literalidade? Uma reiteração da consulta preliminar" In: *Ipsis Litteris. Boletim da Associação Profissional dos Tradutores Públicos e Intérpretes Comerciais do Estado de São Paulo*, São Paulo, year 3, n. 11, p. 3.

Hausmann, F. J. 1985. Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz and J. Mugdan (eds.) *Lexikographie und Grammatik*. Tübingen: Niemeyer.

L'Homme, M. and Bertrand, C. 2000. "Specialized lexical combinations: should they be described as collocations or in terms of selectional restrictions?" *Proceedings Ninth Euralex International Congress*, 497-506.

Mayoral-Asensio, R. 2003. *Translation practices explained*. Manchester: St. Jerome Publishing.

O'Keeffe, A.; McCarthy, M. (eds.). 2010. *The Routledge Handbook of Corpus Linguistics*. London/New York: Routledge.

Orenha-Ottaiano, A. 2009. *Unidades fraseológicas especializadas: colocações e colocações estendidas em contratos sociais e estatutos sociais traduzidos no modo juramentado e não-juramentado*. Unpublished Ph.D. Thesis, Universidade Estadual Paulista "Júlio de Mesquita Filho", Brazil.

Scott, M. (2008), *WordSmith Tools*, version 5.0. Liverpool: Lexical Analysis Software Ltd.

Repetition and self-correction in students' interpreting performance: Corpus evidence of the "why" and "how"

Jun Pan
Hang Seng Management College, Hong Kong
janicepan@hsmc.edu.hk

Factors related to the delivery or presentation of interpreting outputs have long been regarded as important in interpreting quality assessment (e.g., Shlesinger 1994; Mead 2000, 2005; Tissi 2000; Ahrens 2005; Pradas Macías, 2006; Rennert 2010). As stated by Gile (2009), presentational factors form a substantial part of the first impression that people render to a communicative act of interpreting or translation:

> Good voice and pleasant delivery, pleasant style and good layout of a printed page can occasionally do more toward convincing a listener or reader than the quality of the idea that is formulated or the information that is delivered. Conversely, good content is weakened by poor style in writing, unusual or inaccurate terminology, a poor voice or poor delivery of a speech. (p. 38)

A growing number of studies have been developed recently to investigate the specific influences of disfluency factors such as pauses and self-repairs in the assessment of professional interpreters' performance (e.g. Tisse 2000; Mead 2005; Pradas Macías 2006); some even involve the analysis of large-scale data in a corpus (e.g., Bendazzoli et al. 2011). These studies provided a lot of useful information about the role of dysfluencies in interpreting quality evaluation and their underlying causes.

Despite the fact that disfluencies occur frequently in student interpretations and are therefore usually included in classroom evaluation schemes (Yang 2005; Cai 2007), there are few studies exploring the "why" and "how" of these problems in students' interpreting performance. Nevertheless, the recognition of problems related to the delivery in interpreting will be beneficial to students' interpreting performance. For example, it is noted by Bartłomiejczyk (2007) that learners' perceptions about presentation problems, if any, could be most effectively translated into enhancement of their actual performance but not perceptions of other problems. Therefore, investigations of students' interpreting delivery through large-scale corpus data will provide significant insights into the "blackbox" of the learning of interpreting and help enhance

greatly the effectiveness and efficiency of interpreter training.

The present study looks into the problems of repetition and self-correction in students' interpreting performance. The study aimed to explore into the "why" and "how" of these problems through the application of corpus analysis methods. A small corpus composed of university students' consecutive interpreting test outputs (Chinese-English consecutive interpreting) was constructed. The corpus included audio files lasting a total of 92,400 seconds (i.e., 1,540 minutes) and their written transcriptions. The audio files were transcribed into computer readable formats to be processed by corpus analysis tools such as Wordsmith 6.0. To fulfil the specific purpose of this study, unique features of spoken text such as pauses, vocalized non-lexical phenomena (e.g. coughs, laughs, etc.), as well as shifts or changes in vocal quality (e.g. change to a soft voice, a possible indication of lack of confidence) were included in the transcription, following the TEI conventions (Sperberg-McQueen and Burnard 2004). The transcription also included features such as pause fillers, silent pauses, small voice, indistinguishable words, extra-linguistic information and errors such as grammar mistakes and pronunciation errors following certain formats (Pan and Yan 2012). Metadata were later added to the transcribed data. In addition, the problems of repetition and self-correction were particularly annotated in this study.

A few previous studies were compared (e.g., Tissi 2000; Mead 2005), and the annotation scheme was finally adapted from that used by Bendazzoli et al. (2011). Bendazzoli et al. (2011), although about simultaneous interpreting, provided the only scheme pertaining to the study of both repetition and self-repair in a corpus-based study. Disfluencies investigated in their study included two sub-categories, i.e., mispronounced words (repetitions) and truncated words. Since the scheme was for the purpose of studying professional interpreters' performance in simultaneous interpreting involving mainly European languages, adaptations were made for its application in the present study. For example, unnecessary subtypes were excluded or merged (e.g. subtypes of the original speech errors including phonological anticipation, phonological perseveration and approximation were combined into the phonological level errors) and new types were added to the current scheme (e.g. the adding of a new subtype of syntactical level speech errors). Although the category of "other" was originally kept in the annotation scheme, it was found that no extra subtypes could fall into this category.

Findings regarding patterns of students' repetition and self-correction problems in consecutive interpreting, their possible causes and features will be reported in this study. The differences between students' performance and that of the professionals will be compared. The pedagogical implications of these findings will also be discussed.

The study will shed important lights on the construction and application of the interpreting learner corpus. It will also provide significant insights into curriculum development and pedagogical enhancement in interpreter training at different levels.

References

Ahrens, B. 2005. "Analysing prosody in simultaneous interpreting: Difficulties and possible solutions". *The Interpreters' Newsletter* 13: 1-14.

Bartłomiejczyk, M. 2007. "Interpreting quality as perceived by trainee interpreters: Self-evaluation". *The Interpreter and Translator Trainer* 1 (2): 247-267.

Bendazzoli, C., Sandrelli, A. and Russo, M. 2011. "Disfluencies in simultaneous interpreting: A corpus-based analysis". In A. Kruger, K. Wallmach, and J. Munday (eds.) *Corpus-based translation studies: Research and applications*. London/New York: Continuum.

Cai, X. 2007. *Kouyi pinggu [Interpretation and evaluation]*. Beijing: Zhongguo Duiwai Fanyi Chuban Gongsi [China Translation and Publishing Company].

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

Mead, P. 2000. "Control of pauses by trainee interpreters in their A and B languages". *The Interpreters' Newsletter* 10: 89-102.

Mead, P. 2005. "Methodological issues in the study of interpreters' fluency". *The Interpreters' Newsletter* 13: 39-63.

Pan, J. and Yan, J. X. 2012. "Learner variables and problems perceived by students: An investigation of a college interpreting program in China". *Perspectives: Studies in Translatology* 20 (2): 199-218.

Pradas Macías, M. 2006. "Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency". *Interpreting* 8 (1): 25-43.

Rennert, S. 2010. "The impact of fluency on the subjective assessment of interpreting quality". *The Interpreters' Newsletter* 15: 101-115.

Sperberg-McQueen, C. M. and Burnard, L. 2004. *Text encoding initiative: The XML version of the TEI guidelines*. Available online at http://www.tei-c.org/release/doc/tei-p4-doc/html/

Shlesinger, M. 1994. "Intonation in the production and perception of simultaneous interpretation". In S. Lambert and B. Moser-Mercer (eds.) *Bridging the gap: Empirical research in simultaneous interpretation.*

Amsterdam: John Benjamins.

Tissi, B. 2000. "Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis". *The Interpreters' Newsletter* 10: 103-127.

Yang, C. S. 2005. *Kouyi jiaoxue yanjiu: Lilun yu shijian [Interpretation Teaching and Research: Theory and Practice]*. Beijing: Zhongguo Duiwai Fanyi Chuban Gongsi [China Translation and Publishing Company].

Comparing focus constructions in Brazilian Portuguese and Madrid Spanish

Paulo Pinheiro-Correa
Universidade Federal Fluminense, Capes
papicorrea@gmail.com

1 Overview

This paper presents the first results of a postdoctoral research in progress on the realization of the pragmatic function *focus* in Brazilian Portuguese and Madrid Spanish. We are analyzing data from two comparable corpora: *C-Oral Rom* (Spanish) and *C-Oral Brasil*, two corpora with the same kind of segmentation, phonetically-based. We aim to describe the equivalences of the different kinds of focus in both languages. Martínez Caro (1995), comparing the realization of focus constructions in spoken Madrid Spanish and London English shows that while Spanish tends to mark different focuses syntactically, intonation plays a significant role in the marking of focuses in English, confirming Lambrecht's 1996 statements on the difference between focus realization across languages. Following Martínez Caro's study we aim to describe the possibilities of narrow focus realization in Brazilian Portuguese and Spanish.

2 Hypothesis

Our hypothesis is that corpus study could reveal the possibility of narrow (contrastive) focus marking in Brazilian Portuguese purely by means of intonation, as it is informed for English, instead of a combining syntactic and prosodic or a purely syntactic marking, already described for it. There could be a syntactic reason for a purely prosodic marking of narrow focus in this romance language. Kato 1999 among other authors consider BP a language in process of linguistic change, switching from a null subject parameter such as Spanish towards a full subject language, such as English. The shift to a non-null subject parameter has many other syntactic effects such as word order tending to be more fixed as the new parameter is set.

A previous corpus-based study, comparing a same TV show produced in Brazil and Argentina (Moura 2013) showed that while in Argentine Spanish several types of narrow focuses were marked through a different word order or by means of cleft sentences, Brazilian Portuguese data – besides showing a wide range of cleft and pseudo-cleft sentences marking narrow focus – showed also the possibility of purely intonational marking of narrow

focus, a feature we are dealing with in this paper.

As Brazilian Portuguese could be thought as a language in which there are two competing grammars, a conservative one and an innovative one, the possibility of there being a purely intonational marking of narrow focus could be related to the latter.

3 Methodology

To investigate such features we used *Praat, version* 62, an acoustic analysis software, since both C-Oral corpora we are dealing with present all the records of the data and we combine an acoustic analysis, a phonological marking of prosodic features (with ToBI notation) and a syntactic one, based on a set of cleft and pseudo-cleft constructions conveying focus in Brazilian Portuguese (Braga et al 2009). We adapted this classification to other possibilities of focusing in both languages, such as word order changes and the presence vs absence of subject pronouns for Spanish and the difference between weak pronominals and strong pronouns in Brazilian Portuguese, in order to get a syntactic and acoustic description of narrow focuses in Brazilian Portuguese (Belo Horizonte) and Spanish (Madrid).

References

Braga, M. L., Kato, M.A and Mioto, C. 2009. "As construções-Q no português brasileiro falado." In M. Kato and M. do Nascimento (eds.). *Gramática do português culto falado no Brasil – A construção da sentença*. Campinas: Editora da Unicamp.

Boersma, P and D. Weenink. *Praat*.

Cresti, E. and M. Moneglia (eds.). 2005. *C-Oral Rom*.

Lambrecht, K. 1996. *Information Structure and Sentence Form. Topic, Focus, And The Mental Representations Of Discourse Referents*. Cambridge, UK, Cambridge University Press.

Kato, M. 1999. "Strong pronouns, weak pronominals and the null subject parameter." *Probus* 11,1. 1-37.

Martínez Caro, E. 1995. "*Funciones pragmáticas, orden de constituyentes y acentuación en inglés y en español. Estudio de corpus." PhD Thesis. Universidad Complutense de Madrid.*

Moura, F. C. S. 2013. "A função informativa foco em um estudo comparativo português-espanhol." *M.A. Dissertation, Universidade Federal Fluminense.*

Raso, T. and H. Mello. 2012. *C-Oral Brasil 1*.

Informational load as a trigger for disfluencies in interpreting: A corpus-based regression analysis

Koen Plevoets **Bart Defrancq**

Ghent University

koen.plevoets@ugent.be bart.defrancq@ugent.be

One of the major aspects of an interpreting task is the high cognitive load for the interpreter. Gile (1995) pinpoints the interpreter's lack over the conceptual content and his reduced background knowledge (in comparison to the speaker) as potential sources of problems during interpreting, and for simultaneous interpreting he lists the additional obstacles of the lack of control over the original speech rate and the mutually detrimental influence of the speaking and listening task. Psycholinguistic research (Clark et al. 2002; Corley et al. 2008; Watanabe et al. 2008) has revealed that information overload is prone to give rise to disfluencies in the utterance, e.g. *uh* or *uhm*. In light thereof, it is no surprise that disfluencies figure prominently in interpreting (Bakti 2009, Tissi 2000, Tóth 2011). However, previous research is inconclusive as to whether disfluencies occur to the same extent in interpreting as in spontaneous speech, as no systematic quantitative comparison has yet been undertaken.

This paper will analyse the relation between interpreting, informational load and disfluencies in a corpus of interpreted language as compared to a corpus of spontaneous speech. The corpus of interpreted language was compiled at Ghent University between 2010 and 2013. It consists of French, Spanish and Dutch interpreted speeches in the European Parliament from 2006 until 2008. The audio fragments were transcribed according to the guidelines of the VALIBEL corpus (Bachy et al. 2007). For our purposes, a sub-corpus of French source speeches and their Dutch interpretations was selected, amounting to a total corpus size of 140 000 words. The sub-corpus has additionally been annotated for lemmas, parts-of-speech and chunks (Van de Kauter et al. 2013). The corpus which serves as the reference for spontaneous speech is the sub-corpus of political debates of the Spoken Dutch Corpus (Oostdijk 2000). This sub-corpus contains 220 000 words of Netherlandic Dutch and 140 000 words of Belgian Dutch, which were collected between 1998 and 2003, and it is annotated for lemmas and parts-of speech.

In both corpora, each sentence (or 'discourse unit') was subsequently coded for informational measures such as lexical density and syntactic depth,

in order to capture the informational load experienced by the speakers or interpreters. The measurement of lexical density is based on the POS-tags, where all nouns, non-auxiliary verbs, adjectives and adverbs are counted as content words and all pronouns, auxiliary verbs, prepositions, conjunctions and determiners are counted as function words (all remaining interjections and fillers are treated as a rest category). The coding for syntactic depth was done manually: each sentence was screened and annotated for the number of different syntactic subordinations, the maximal degree of syntactic subordination and the average degree of syntactic subordination. The last step in the data retrieval consisted in counting the number of the disfluencies *uh* and *uhm* per sentence, as the aim of the analysis is to predict the frequency of the disfluencies on the basis of the informational load of each sentence. Due to the heavy skewness of the frequency data, it was decided to run the analysis by means of Robust Regression (Maronna et al. 2006).

The results confirm the intuitive assumptions in that the data for interpreted Dutch exhibit a different pattern from both the data of the French source language and of spontaneous Dutch, which in turn are very similar. The observations for interpreted Dutch show a distinctly positive effect of the informational measures on the frequency of the disfluencies: the higher the informational load is during interpreting, the more this results in disfluencies by the interpreter. A striking finding for both the French source data and the spontaneous Dutch data is that the effect in either case is negative. This result may be attributed to the highly prepared nature of the parliamentary speeches, which are sometimes read out verbatim from a written text. The same patterns moreover show up in separate analyses for *uh* and *uhm*. All these findings point to interesting prospects for further research. The immediate next step will be to take account of the position of the disfluency in the utterance, as we conjecture that disfluencies tend to occur before informationally heavy chunks in non-interpreted language, but at the onset of whole utterances in interpreted language.

References

Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.C. and Wilmet, R. 2007. *Conventions de transcription régissant les corpus de la banque de données VALIBEL*. University of Louvain-la-Neuve. Available online at http://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.PDF.

Bakti, M. 2009. "Speech disfluencies in simultaneous interpreting". In D. De Crom (ed.) *(Trans)formation of identities: Selected papers of the CETRA research seminar in translation studies 2008*. Leuven: CETRA, 1-18.

Clark, H.H. and Fox Tree, J.E. 2002. "Using *uh* and *um* in spontaneous speaking". *Cognition* 84: 73-111.

Corley, M. and Stewart, O.W. 2008. "Hesitation disfluencies in spontaneous speech: The meaning of *um*". *Language and Linguistics Compass* 2: 589-602.

Gile, D. 1995. *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.

Maronna, R., Martin, D. and Yohai, V. 2006. *Robust statistics: Theory and methods*. Hoboken, New Jersey: John Wiley and Sons.

Oostdijk, N. 2000. "The Spoken Dutch Corpus: Overview and first evaluation". *Proceedings of the Second International Conference on Language Resources and Evaluation*: 887-894.

Tissi, B. 2000. "Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis". *The Interpreters' Newsletter* 10: 103-127.

Tóth, A. 2011. "Speech disfluencies in simultaneous interpreting: A mirror on cognitive processes". *SKASE Journal of Translation and Interpretation* 5: 23-31.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L. and Hoste, V. 2013. "LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit". *Computational Linguistics in the Netherlands Journal* 3: 103-120.

Watanabe, M., Hirose, K. Den, Y. and Minematsu, N. 2008. "Filled pauses as cues to the complexity of up-coming phrases for native and non-native listeners". *Speech Communications* 50: 81-94.

Contrasting impersonal strategies in English and Russian on the basis of translational corpora

Olga Rudolf
Friedrich-Schiller University Jena
olga.rudolf@uni-jena.de

This corpus-based study focuses on the comparison of human impersonal strategies in English and Russian, i.e. pronouns and constructions that generalize over a set of human individuals as in (1a, b).

(1) a. ***One*** *should never lie*.
b. (Russian)
V Germanii ***ljubjat*** *pit' pivo*.
in Germany like.3PL.PRS drink beer
"They like drinking beer in Germany."
c. (German)
Man sollte zu Älteren höflich sein.
man should.3SG to elder polite be
"One should be polite to elder people."

Some languages have dedicated impersonal pronouns, e.g. *man* in German (as in 1c), which can be used to express distinct interpretations of the human referent (see Dimova 1981; Zifonun 2001), whereas English and Russian have no such specialized expression and use other means to express impersonalisation. English typically uses the personal pronouns *you*, *they* or *one* with impersonal meanings, and in Russian impersonalisation is basically encoded in verbs (3rd-person plural and 2nd-person singular or modal impersonal) with null subject pronouns. These and other impersonal constructions in English and Russian differ from one another in a number of features, most importantly (i) the interpretation of the human referent, which can be described in terms of quantification (universal/existential) and (internal/external) perspective, i.e. whether or not the speaker identifies him/herself with the referent, and (ii) the contexts where the constructions occur, such as generalizing/episodic situations and veridical/non-veridical propositions.

According to claims made in the literature, which say that the distinct readings of the human referent are triggered by their sentential context, e.g. insofar as generic contexts give rise to a universal interpretation (see Moltmann 2010; Malamud 2006), the corpus study was, in a first step, aimed at testing correlations between the types of context and the readings of the referent. Two separate corpus studies were conducted, using data from German-English and German-Russian parallel corpora (ParaSol for both studies). German was taken as a basis for the studies because its pronoun *man*, unlike most English and Russian impersonal constructions, is always unambiguously impersonal and can be conveniently searched for in the corpus. This way translation strategies of German *man* and inventories of impersonal constructions in English and Russian can be identified.

The dichotomies characterizing interpretations of the referent and contexts mentioned above were used as binary variables to code the data and were then checked for correlations. Some statistical tests showed that there is indeed a strong correlation between the context variable *generalizing* and a *universal* interpretation of the referent with *internal* perspective. On the other hand, the context variable *veridical* does not correlate with the other variables but has a significant influence on the choice of the translation strategies in both English and Russian. Therefore, in a next step, the combination of the first three variables on the one hand, and the variable *veridical* on the other, were taken as independent parameters in multinomial logistic regression analyses in order to make predictions about the choice of translation strategies found in English and Russian. The results present a hierarchy of probabilities predicting the occurrence of each strategy under specific conditions. For example, if the context is veridical, the most likely strategy to occur is the 3rd-person plural in Russian, the second likely one is a 2nd-singular form, and the least likely strategy is a modal impersonal.

Generalizations about the occurrence of impersonal strategies in English and Russian under certain semantic conditions allow for comparing/contrasting selected strategies in these languages. The (in terms of frequency) major strategies in the two languages constitute partial equivalents, e.g. the English pronoun *they* and the 3rd-plural form in Russian, or English *you* and Russian 2nd-singular forms. In spite of their quite different grammatical forms, they show largely similar behaviour from a functional perspective. For example, both English *they* and Russian 3rd-plurals tend to occur in veridical and episodic sentences with an existential referent and external perspective, though they are not fully equivalent: the Russian construction can also be used to render an internal perspective, which is impossible in English. English *you* and Russian 2nd-singular forms always occur in generalizing sentences and can only take an internal perspective. The English pronoun *one* does not have an equivalent in Russian, and only Russian, in turn, has a modal impersonal strategy. These two strategies, however, seem to constitute a functional pair, since both are mainly used in modal contexts with universal referents and internal perspective.

References

Dimova, A. 1981. „Die Polysemie des Pronomens *man* in der deutschen Gegenwartssprache und die Kontextbedingungen für seine Monosemierung.“ *Beiträge zur Erforschung der deutschen Sprache* 1: 47-75.

Malamud, S. A. 2006. *Semantics and pragmatics of arbitrariness*. PhD. U of Pensylvania.

Moltmann, F. 2010. “Generalizing Detached Self-Reference and the Semantics of Generic *One*”. *Mind & Language* 25: 440–473.

Zifonun, G. 2001. „Man lebt nur einmal. Morphosyntax und Semantik des Pronomens *man*". *Deutsche Sprache* 28: 232–253

Corpus Linguistics and Translation Studies: a study of corpus teaching methodologies applied to the reading of the Italian translations of Joyce's *A Portrait of the Artist as a Young Man*

Chiara Sciarrino
University of Palermo
chiara.sciarrino@unipa.it

The use of Corpus Linguistics within a language classroom has undergone a considerable increase during the last years. Corpus linguistics and stylistics are also fairly extensively used within the field of translation studies, as shown by Laviosa (2002), amongst others. Corpora can indeed disclose features of the translated texts, detect all those stylistic characteristics that are typical of a text and its translation into another language as well as help in the teaching of translation practice or in the teaching of a foreign language. Corpus linguistic techniques are used by Johnson (2007) to investigate what stylistic features emerged from multi wordlists of a corpus of works by the Italian writer Grazia Deledda. Findings were later compared to the English translations of her novels with the final aim to give suggestions on better ways of translating. Also, she suggested that 'a corpus stylistic approach could also be exploited by literary translators in order to begin the task of translation with a more thorough knowledge of the Source Text'. Johnson's claim that 'It would also be feasible to use a corpus stylistics approach descriptively to evaluate the success of a particular translation or compare different translations of the same text' is here taken as a starting point for the current investigation.

This paper explores the impact that the availability of the techniques and tools of corpus linguistics is likely to have on the study of literary translation. In particular, the linguistic analysis of translation corpora of the Italian editions of James Joyce's *A Portrait of the Artist as a young Man* is here undertaken with the aim of evaluating the translation process itself. What happens in the process of translation and what are the results obtained throughout time by the different translators are some of the issues that will also be considered.

The use of an electronic corpus in a postgraduate course on 'English language and translation studies held at the University of Palermo was introduced with the aim of providing empirical data and authentic material alongside the actual copies of the translations themselves. Given the short amount of hours at disposal, it was necessary to have an instrument which could quickly provide students

with both quantity and quality data in a relatively short period of time.

Each student was assigned a specific passage from the original text and was asked to compare the Italian translations so far published. In particular, passages describing reflections about religion and faith, which dominate chapter III of the novel, were chosen and a specific objective was set: to analyse the semantic area of religion. Differences between the various ways in which religious terms have been rendered into Italian were stressed by students. If from one hand their 'manual' work through the pages of the book highlighted the presence of some collocations within specific contexts, from the other hand, the research was made more 'visible' through the use of special software like *Paraconc* and *Wordsmith*, which helped students better memorize specific narrative and linguistic information not easily detectable.

References

Baker, M. (1993), 'Corpus linguistics and Translation Studies: Implications and applications', in *Text and Technology: in Honour of John Sinclair*, ed. by M. Baker et al, John Benjamins, Amsterdam/Philadelphia, pp. 233-250.

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

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*, ed. by H. Somers, John Benjamins, Amsterdam/Philadelphia, pp. 175-186.

Bernardini, S. and F. Zanettin, (ed. by), (2000) *I corpora nella didattica della traduzione. Corpus use and learning to translate.* Atti del Seminario di Studi Internazionale Bertinoro 14-15 novembre 1997, Proceedings of the International Workshop Bertinoro 14-15 November 1997, Bologna, CLUEB.

Granger S., J. Lerot and Petch-Tyson S. (2003), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Rodopi, Amsterdam, New York.

Jantunen, J. H. (2002), 'Comparable Corpora in Translation Studies: Strengths and Limitations', in *Sky Journal of Linguistics*, 15: 105-117.

Johnson, J. H. (2010), 'A Corpus-assisted Study of *PARERE/SEMBRARE* in Grazia Deledda's Canne al Vento and La Madre. Constructing Point of View in the Source Texts and their English Translations', in Douthwaite J., Wales K. (eds.) (2010), *Stylistics & Co. (Unlimited). The Range, Methods and Applications of Stylistics*, Textus. English Studies in Italy, vol. XXIII, no. 1 (January-April), pp. 283-302.

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

Munday, J., 'A Computer-Assisted Approach to the Analysis of Translation Shifts' in *Meta*, XLIII, 4, 1998.

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

Zanettin, F. (2000), 'Parallel Corpora in Translation Studies: Issues in Corpus Design', in *Intercultural Faultlines* in Olohan, M. (ed. by), *Research Methods in Translation studies*, St. Jerome, Manchester, pp. 105-118.

Sentence splitting in the translation pair English-German

Tatiana Serbina
RWTH Aachen
serbina@anglistik.rwth-aachen.de

Originals and the corresponding translations are often characterized by a number of translation shifts that can be identified between linguistic units of various sizes. Recently, translation shifts have been investigated using corpus methods (e.g. Čulo et al. 2008). One type of such shifts is sentence splitting, which is in the focus of the present corpus-based study: this phenomenon takes place when the sentence boundaries are shifted by translating one source sentence by two or more sentences (Ramm 2004).

Previous research has indicated that in the process of translation source text structures such as coordinated and subordinated clauses, as well as complex noun phrases can be separated into independent sentences (Fabricius-Hansen 1999, Ramm 2004, 2006, Solfjeld 2008). The occurrences of this type of translation shifts have been mainly explained through a number of contrastive differences. These could be, for instance, differences in noun phrase modification. Most of the previous studies have concentrated on the language pair German-Norwegian and performed mainly qualitative analyses: while some quantitative information is included, it is not submitted to statistical testing. Therefore, one aim of this study is to analyse in a quantitative manner whether the same grammatical structures trigger sentence splitting in translations from English into German, and the opposite translation direction considering the relevant contrastive differences. Moreover, an additional explanatory factor, namely register, is taken into account.

Sentence splitting is said to reduce information density by distributing the information across several target sentences. Since not only simple, but also complex sentences are assumed to be processed as whole units (Fabricius-Hansen 1999), it is possible that translators split sentences due to high processing demands: several shorter sentences could be easier to process than one complex sentence with high information density. The phenomenon of sentence splitting could also function as a conscious translation strategy to simplify the target text for the reader, even though this might not always have the desired effect (cf. Wolfer et al. 2013). Instances of sentence splitting, especially when a phrase in the original corresponds to a sentence in the translation, have been also interpreted in terms of the translation property of explicitation (Fabricius-Hansen 1999).

The present study uses the CroCo corpus, a parallel corpus compiled for the language pair English-German. The corpus contains approximately one million words and is subdivided into eight registers, namely political essays (ESSAY), fictional texts (FICTION), instruction manuals (INSTR), popular scientific texts (POPSCI), letters to shareholders (SHARE), prepared speeches (SPEECH), tourist leaflets (TOU) and webpages (WEB). Its multi-level annotation and alignment allows querying for and extracting of translation shifts realized through the so-called crossing lines, for instance between clauses and sentences: in these cases the aligned clauses expressing the same semantic information belong to different sentences (Hansen-Schirra et al. 2012).

A comparison of the number of sentences in English originals and German translations indicates that in six out of eight registers instances of sentence splitting could be expected: with the exception of the registers FICTION and SHARE there are more sentences in German translations than in English originals. In this translation direction it is especially the register POPSCI that is characterized by an increased number of sentences. In contrast, the English translations from German contain fewer sentences than the corresponding originals, irrespective of the register.

More detailed analyses of the examples containing the investigated phenomenon are required. Thus, it should be taken into account that sentence boundaries can be also changed in a variety of other ways: several sentences in the original can be merged into one, parts of source sentences can be attached to other sentences, and the whole sentences can be missing (Ramm 2004) or be added in the translation. These shifts certainly affect the number of sentences in originals and translations as well.

The present study scrutinizes individual cases of sentence splitting belonging to different registers to gain more insights into the nature of this phenomenon. A quantitative investigation compares the relative contribution of various triggers to the overall number of sentences split in translation. Moreover, it is also analysed how often these grammatical structures shift or are kept in translations. The results of the study will further our understanding of translation shifts leading to possible applications in machine translation or teaching of translation.

References

Čulo, O., Hansen-Schirra, S., Neumann, S. and Vela, M. 2008. "Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus". In *Proceedings of the LREC 2008 Workshop*

'Building and Using Comparable Corpora'. Marrakesh, Morrocco, 47–51.

Hansen-Schirra, S., Neumann, S. and Steiner, E. 2012. *Cross linguistic corpora for the study of translations: insights from the language pair English-German*. Berlin: de Gruyter.

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

Ramm, W. 2004. "Sentence-boundary adjustment in Norwegian-German and German-Norwegian translations: first results of a corpus-based study". In K. Aijmer and H. Hasselgard (eds.) *Translation and Corpora*. Gothemborg: Acta Universitatis Gothoburgensis, 129-147.

Ramm, W. 2006. "Dispensing with subordination in translation: consequences on discourse structure". In T. Solstad, A. Grønn and D. Haug (eds.) *A Festschrift for Kjell Johan Sæbø: in partial fulfilment of the requirements for the celebration of his 50th birthday*. Oslo: Oslo University, 121-136.

Solfjeld, K. 2008. "Sentence splitting – and strategies to preserve discourse structure in German-Norwegian translations". In C. Fabricius-Hansen and W. Ramm (eds.) *'Subordination' versus 'Coordination' in sentence and text: a cross-linguistic perspective*. Amsterdam: Benjamins, 115–133.

Wolfer, S., Hansen, S. and Konieczny, L. 2013. "Are shorter sentences simpler? Discourse level processing consequences of reformulating texts". 7th EST Congress. Germersheim, Germany.

Modal and post-modal uses of Lithuanian adverbials: evidence from a parallel corpus

Audronė Šolienė
Vilnius University
audrone.soliene@gmail.com

1 Introduction

Contrastive studies based on parallel and comparable corpus data (Aijmer 1996, 1999; Johansson 2001, 2007; Simon-Vandenbergen and Aijmer 2007; Mortelmans 2010 among others) show that in a cross-linguistic perspective the degree of lexical correspondence in expressions of epistemic modality is not very high and different subsystems tend to interact. This phenomenon is explained in terms of structural cross-linguistic differences as well as different degrees of grammaticalization, pragmaticalization and/or polyfunctionality of modal markers.

Polyfunctionality is a common phenomenon in many languages. Great attention has been paid to modal verbs (auxiliaries) and their epistemic, deontic and dynamic interpretation in different languages (Coates 1983; Hoye 1997; Palmer 2001; Holvoet 2009 and others). Adjectives can also have epistemic or dynamic readings (Lyons 1977). Recent research has shown that epistemic modal adverbs can be used in different ways as well (Simon-Vandenbergen and Aijmer 2007; Pietrandrea 2008; Cornillie 2010). Modal adverbs do not usually convey dynamic or deontic readings; however, besides their epistemic status, they can have a variety of slightly different, post-modal, interpretations, e.g.:

(1) *Could you perhaps explain it?*

Lithuanian modal adverbials have not yet been looked at in great detail, nor have they been explicitly compared with their English correspondences in terms of polyfunctionality. As no consensus has been reached so far regarding the distinction between the word classes of modal particles and adverbs in Lithuanian linguistics, the term 'adverbials' is used to cover both (Smetona and Usonienė 2012). The present paper aims to investigate the modal and post-modal uses of Lithuanian polyfunctional adverbials *gal* 'perhaps' and *galbūt* 'maybe': to determine their functional variants in different discourse types and to establish parallels between the function and form with the help of the analysis of their translational correspondences.

2 Data and methods

The corpus-based approach adopted in this study helps to reveal patterns and meanings of modal expressions which would be difficult to find otherwise. The method used in the research is non-experimental data collection; it is a contrastive analysis based on the data obtained from a self-compiled bidirectional parallel corpus – *ParaCorp*$_{EN→LT→EN}$ (Šolienė 2013). The corpus design follows the model of the English-Norwegian Parallel Corpus (Johansson 2007). The *ParaCorp*$_{EN→LT→EN}$ was compiled from original English fiction texts and their translations into Lithuanian and original Lithuanian fiction texts and their translations into English. The size of the corpus is about 5M words.

Since the sub-corpora are of different size, the raw frequency numbers have been normalized per 10, 000 words. Furthermore, in order to check whether the similarities and differences are statistically significant, the log-likelihood test was performed, which is commonly considered to be a more statistically reliable test than the chi-square test (cf. Dunning 1993). Frequencies of particular patterns and uses are of crucial importance to this study, since frequency can be an important factor in specification of meaning (Leech 2003; Simon-Vandenbergen and Aijmer 2007). Some of the tendencies identified in the parallel corpus were verified in other databases: the Corpus of the Contemporary Lithuanian Language [20] and the Corpus of Academic Lithuanian[21].

3 Results and preliminary observations

The investigated adverbials *gal* 'perhaps' and *galbūt* 'maybe' as well as their English counterparts mainly serve as markers of epistemic modal possibility, which is attributed to them as their main function by various dictionaries and grammars. Though the adverbial *gal* 'perhaps' is more versatile in terms of polyfunctionality, it is clear that both adverbials have developed post-modal uses. The markers exhibit a diversity of functional variants in different types of discourse: they can act as intensifiers of the alternative, which emphasizes the choice between several options; as mitigating devices reducing the illocutionary effect of an utterance; as interrogative particles; as approximators estimating a figure, number or quantity.

References

Aijmer, K. 1996. "Swedish modal particles in a contrastive perspective". *Language Sciences* 18: 393–427.

Aijmer, K. 1999. "Epistemic possibility in an English-Swedish contrastive perspective". In H. Hasselgård and S. Oksefjell (eds.) *Out of corpora. Studies in honour of Stig Johanson*. Amsterdam: Rodopi. 301–321.

Coates, J. 1983. *The Semantics of Modal Auxiliaries*. London: Croom Helm.

Cornillie, B. 2010. "An Interactional Approach to Evidential and Epistemic Adverbs in Spanish Conversation". In G. Diewald and E. Smirnova (eds.) *Linguistic realization of evidentiality in European Languages*. Berlin & New York: Mouton de Gruyter.

Dunning, T. 1993. "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics* 19 (1): 61–74.

Holvoet, A. 2009. "Modals in Baltic". In B. Hansen and F. de Haan (eds.) *Modals in the languages of Europe. A reference work*. Berlin: Mouton de Gruyter. 199-228.

Hoye, L. 1997. *Adverbs and Modality in English*. London & New York: Longman.

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

Johansson, S. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies. (Studies in Corpus Linguistics, 26)*. Amsterdam & Philadelphia: John Benjamins.

Leech, G. 2003. "Modality on the Move: The English Modal Auxiliaries 1961-1992". R. Facchinetti, M. Krug and F. Palmer (eds). *Modality in Contemporary English*. Berlin: Mouton de Gruyter. 191–240.

Lyons, J. 1977. *Semantics*. Cambridge University Press.

Mortelmans, T. 2010. "Epistemic must and its Cognates in German and Dutch. The Subtle Differences". Unpublished manuscript. University of Antwerp.

Palmer, F.R. 2001. *Mood and Modality* [2nd ed.]. Cambridge: Cambridge University Press.

Pietrandrea, P. 2008. "Certamente and sicuramente: Encoding dynamic and discursive aspects in Italian". *Belgian Journal of Linguistics* 22: 221–246.

Simon-Vandenbergen, A.M. and Aijmer, K.. 2007. *The semantic field of modal certainty: a corpus-based study of English adverbs*. Berlin & New York: Mouton de Gruyter.

Smetona, A. and A. Usonienė. 2012. "Autoriaus pozicijos adverbialai ir adverbializacija lietuvių mokslo kalboje". *Kalbotyra* 64 (3): 124–139.

Šolienė, A. 2013. *Episteminio modalumo ekvivalentiškumo parametrai anglų ir lietuvių kalbose*. Unpublished PhD thesis, Vilnius University. Available online at http://vddb.laba.lt/fedora/get/LT-eLABa-0001:E.02~2013~D_20130701_092318-53000/DS.005.0.01.ETD.

[20] http://donelaitis.vdu.lt
[21] http://www.coralit.lt

A parallel corpus-based investigation of vocabulary features of tourism translations

Sun Yifeng
Lingnan University
sunyf@ln.edu.hk

Tang Fang
Guangdong University of Foreign Studies
candy.fangtang@hotmail.com

1 Introduction

This paper explores the vocabulary features in tourism translations from Chinese into English. A paralleled corpus of tourism texts has been built with texts collected from the bilingual tourism websites of Hong Kong, Taiwan, Singapore and the Chinese mainland. All the texts have been POS-tagged (the English version is tagged by TreeTagger and the Chinese version by ICTCLAS 2013) and later manually checked by the two investigators. These texts are also sentence-aligned by using Paraconc software so that they can be searched in a Concordancer alphabetically or retrieved lists of linguistic data based on key words or phrases. The frequencies of each part of speech in the two versions have been compared.

2 Research Questions

1) What is the feature of tourism translations in the usage of nouns and verbs?

2) What is the feature of tourism translations in the usage of superlative forms?

3) What is the feature of tourism translations in the usage of pronouns?

3 Lexical Features

We counted the frequency of nouns and verbs in the two subcorpora. As is shown in Table 1, there are more nouns and fewer verbs in the translated texts, which indicates a possibility of nominalization.

	nouns	verbs
ori	171699	63300
trans	186059	50745

Table 1. Frequency of Nouns and Verbs

In English, verbs can be transformed into nouns by adding suffixes, such as –tion(s), -sion(s), -ment(s), ence(s), -ance(s). The frequency of these suffixes can be seen from Table 2.

	trans
-tion(s)	5517
-sion(s)	330
-ment(s)	958
-ence(s)	709
-ance(s)	943
total	8457

Table 2. Frequency of Nominalization

As proposed by Halliday (1985: 91), nominalization can set writers free from the context and produce a text which is more objective and formal. In the tourism translations we collected, a large amount of nominalization has been identified, which seems to suggest that the translated texts are with higher degree of objectivity than their Chinese originals.

In Chinese, the superlative form of adjective and adverb can be realized simply through the addition of "*zui*" (最, which can roughly be translated as "most" in English). In English, for adjective and adverb with no less than three syllables, this form can be realized through the addition of "most" while for those with less than three syllables, it can be realized by adding the suffix "–est". To investigate the frequency of superlative forms in the two corpora, AntConc 3.3.5 has been adopted by searching "*zui*" in the original corpus and "most" as well as "-est" in the translated corpus. Irrelevant cases like "forest", "destination", etc. have been excluded manually. The result is shown in Table 3, where far more superlative forms can be identified from the translated corpus than the original one. It demonstrates that more superlative forms have been added through translation. It can be regarded as the evidence for translators' emotional involvement and a tendency of intensification of the original meaning. Similar with the previously-mentioned higher frequency in using second-person pronouns, this kind of addition may motivate the readers to a greater extent.

	superlative forms
ori	2303
trans	2605

Table 3. Frequency of superlative forms

The adoption of personal pronouns can usually reflect the writing style. For instance, the frequent use of first-person pronouns indicates the writer's self-centered perspective while the frequent use of third-person pronouns implied a sense of objectivity. According to Reiss (1971), there are three types of texts, namely, information texts, expressive texts and appellative texts. For instance, news is mainly informative, prose is usually expressive and

advertisement is dominantly appellative. Tourism Text can be both informative and appellative. To achieve appellative effects, the writer needs to get the reader involved in what has been described. In this case, the use of second-person pronouns can be regarded as an effective tool to get the writer closer to the readers and can even show a sense of friendliness and hospitality. In this study, AntConc 3.3.5 has been used to find out all the "*ni (men/de)*" (你（们/的）, "*nin (men/de)*" (您（们/的）in the original corpus and "you, your, yours, yourself, yourselves" in the translated corpus. As shown in Table 4, there are more second-person pronouns in the translated tourism texts than their originals, which indicates that the translators have added a large number of second-person pronouns while translating. This act may be conscious or sub-conscious. Yet it can definitely increase the readers' involvement.

	second-person pronouns
ori	409
trans	1566

Table 4. Frequency of Second-Person Pronouns

In this study, AntConc 3.3.5 has been employed to identify the frequency of pronouns in the translated texts as well as their Chinese originals. As is shown in Table 5, statistics reveal that much more pronouns have been adopted in the translation. This suggests that some nouns or nouns phrases in the originals have been replaced by pronouns in the target texts. For instance, words like "visitor" and "traveller" have been replaced by the third-person pronoun "he/ him", which forms a kind of implicitation, namely, "a stylistic translation technique which consists of making what is explicit in the source language implicit in the target language, relying on the context or the situation for conveying the meaning" (Vinay & Darbelnet 1958/1995:344).

	reference
ori	5583
trans	9295

Table 5. Frequency of reference implicitation

4. Conclusion

To sum up, this corpus-based study finds that compared with their Chinese originals, the translated English tourism texts are characterized by: 1) a large amount of nominalized verbs; 2) a higher frequency in the use of second-personal pronouns; 3) a higher frequency in the use of superlative forms; and 4) a higher frequency of pronouns. These features support the "explicitation" and "simplification" with regard to translation as a universal hypothesis. Moreover, this study identifies specific practices of implicitation in the translated texts mainly concerning some historical content that has been deleted. Such implicit stylistic change concerning the linguistic and stylistic features of the final translation product may well be motivated by a strategic consideration of establishing cross-cultural functional equivalence between Chinese and English tourism texts on the part of the translator. In this connection, vocabulary features which contribute to lexical and cross-cultural complexity will also be discussed as manifest in combining continuity and change in the translated texts.

References

Baker, M. 1996. "Corpus-based Translation Studies: the Challenges that lie ahead." In H. Somers (ed.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam: John Benjamins Publishing Company.

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

Baker, M. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Halliday, M. 1985. *Spoken and Written Language*. Victoria: Denkin University Press.

Ji, M. 2010. *Phraseology in Corpus-based Translation Studies*. Berlin: Peter Lang.

Kenny, D. 2001. *Lexis and Creativity in Translation: A Corpus-based Study*. Manchester and Northampton: St. Jerone Publishing.

Kruger, A. J. Munday & K. Wallmach. 2011. *Corpus-based Translation Studies: Research and Applications*. London: Continuum.

Laviosa, S. 1998. "Core patterns of lexical use in a comparable corpus of English narrative prose". *Meta* 43(4): 1-14.

Laviosa, S. 2002. *Corpus-based Translation Studies*. Amsterdam and New York: Rodopi.

Martin, W. 2005. "Stylistics: Corpus Approaches". In K. Brown. (ed.) *The Encyclopaedia of Laanguage and Linguistics*. Oxford: Elsevier.

McEnery, A.M. Tono, Y. and Xiao, Z. 2006. *Corpus Based Language Studies*. London: Routledge.

Olohan, M. 2004. *Introducing Corpora in Translation Studies*. London & New York: Routledge.

Reiss, K. 1971. "Type, kind and individuality of text: decision making in translation". In L. Venuti (eds.) *The Translation Studies Reader*. London: Routeledge.

Scott, M. and Tribble, C. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education*. Amsterdam: John Benjamins Publishing Company.

Semino, E. & M. Short. 2004. *Corpus Stylistics: Speech,*

Writing and Though Presentation in a Corpus of English Writing. London: Routledge.

Vinay, J. & J. Darbelnet. 1958. *Comparative Stylistics of French and English: A Methodology for Translation*. Translated and edited by J. C. Sager & M. J. Hamel. 1995. Amsterdam & Philadelphia: John Benjamins.

Parallel corpora and functionalist-oriented contrastive studies

Beata Trawinski
Institut für Deutsche Sprache, Mannheim
trawinski@ids-mannheim.de

The usefulness of parallel (translation) corpora for cross-linguistic research is widely accepted, and the number of corpus-based cross-linguistic studies is constantly growing. Parallel corpora provide large quantities of bilingual or multilingual aligned authentic language data, offering interesting perspectives for language comparison.

The number of bi- and multilingual corpora is also growing rapidly. Nowadays, there exist parallel corpora for many language pairs, and the technology used for their development is quite advanced. Many of the existing parallel corpora are lemmatized and annotated morphosyntactically, meaning that research questions of the following kind can now be addressed, and the results analyzed quantitatively:

- What are the equivalents of lemma / word form / phrase W in language L1 in languages L2 … Ln?
- Can lemma / word form / phrase W1 in language L1 be expressed by lemma / word form / phrase W2 in language L2?
- How can chains of the grammatical categories C1 … Cn in language L1 be expressed in language L2?
- How can expressions bearing the grammatical features F1 … Fn in language L1 be expressed in language L2?

But research questions of this kind necessarily draw on form-based searches. Semantic queries of the type below cannot be performed using the corpora currently available:

- How is REFERENCE / PREDICATION / QUANTIFICATION / REFLEXIVITY / POSSESSION / PARTITIVITY etc. expressed in languages L1 … Ln?

Precisely this kind of research question, however, has been addressed in our project. The principal goal of our project is to elaborate a grammar of German in comparison with other European languages. The first phase of the project, running from 2001 to 2013, was devoted to the nominal domain. In the second phase, started in 2013, the verbal domain is the subject of investigation. Alongside German, which is the central focus, the core languages for comparison are English, French, Hungarian and Polish, which represent different typological classes. Occasionally, for illustrative or explanatory purposes, other European languages are

consulted, such as Albanian, Basque, Estonian, Finnish, Italian, Dutch, Romanian, Russian, Spanish, Swedish or Turkish.

Unlike the traditional contrastive grammars available for German, which usually cover language pairs, namely German and one another language, and are based on the classical parts of speech and grammatical categories, our grammar is developed rather in the spirit of functionalist typology. This implies that instead of formal criteria, cognitively motivated functional domains are used as a *tertium comparationis.*

This paper discusses the limitations of using parallel corpora in functionalist-oriented contrastive language studies, and presents the conceptual design of a multilingual database of parallel text sequences annotated with functional domains and variance parameters to be compiled in our project.

Using bidirectional parallel corpus data for visualizing differences in semantic structure between translated and non-translated genres. The case of the semantic field of inchoativity in Dutch

Lore Vandevoorde
Ghent University
Lore.Vandevoorde@UGent.be

Gert De Sutter
Ghent University
Gert.DeSutter@UGent.be

Koen Plevoets
Ghent University
Koen.Plevoets@UGent.be

This paper investigates the influence of translation and genre on the structure of semantic fields, thereby tackling the under-researched issue of semantics in Corpus-based Translation Studies. More particularly, it is investigated to which extent the structure of the semantic field of inchoativity differs between original Dutch and translated Dutch, while simultaneously taking into account genre as a potentially influencing variable.

In order to compare semantic fields across genres and varieties, we first have to be able to objectively generate semantic fields for each of the genres and varieties. In this paper, we propose a data-driven, translation-based, bottom-up generation of semantic fields, which is an extension of Dyvik's Semantic Mirroring, a technique for meaning differentiation that uses translational data from parallel corpora. The central idea behind this technique is that translations can be used to identify different senses of a source language word (Dyvik 1998, 2004; Dagan et al., 1991; Lefever, 2012; Aijmer and Simon-Vandenbergen 2004) as well as its lexical relationships. By looking up the translations of an initial lexeme back-and-forth between a source language (under study) and a target language (used as a pivot language), the different meanings of the initial lexeme can be lexicalized, and eventually, visualized via advanced statistical techniques. In this way, an initial lexeme in a language A, e.g. Dutch *bank*, yields translations in a language B, e.g. English *bench, desk, bank* (called *T-image*). When, conversely, the translations of these T-image lexemes are looked up back into Dutch, we end up with an expanded set of lexemes, e.g. Dutch *zitbank* [sofa], *geldbedrijf* [monetary institution], *bank* [bank], *schoolbank* [desk].

Applied to our case study, we first extracted all corpus instances of the Dutch inchoative verb BEGINNEN from the Dutch Parallel Corpus (DPC),

which is both a parallel and comparable corpus of Dutch, French and English (Macken et al., 2011), balanced with respect to five different genres (external communication, journalistic texts, instructive texts, administrative text, fictional and nonfictional literature) and four translation directions (Dutch to French, French to Dutch, Dutch to English and English to Dutch). Second, all French translations of the Dutch lexeme BEGINNEN are checked manually in the DPC (n=292), resulting in a set of 11 unique French translations (the *T-image*). Then, inversely, all translations of the 11 T-image lexemes back into Dutch are looked up (n=823), resulting in 23 unique Dutch lexemes (the *Inverse T-image*). These 23 lexemes are now considered as representative for the semantic field of inchoativity. Finally, the French translations of the Inverse T-image are again queried from the corpus (the *Second T-image*) (n=7079).

We use the (source language) frequencies of the Second T-image and apply the statistical technique of correspondence analysis for visualizing the semantic field of Dutch inchoativity. By doing so, we are able to generate visualizations of the semantic field of BEGINNEN (Figure 1). By using the (target language) frequencies of the Inverse T-image, we can compare visualizations of original (Figure 1) with translated language (Figure 2). Finally, we also generate genre-specific semantic fields for each of the text types available in the corpus.

The visualized results show structural resemblances and small but noteworthy differences between the semantic fields of original texts and translations, as translations seem to flatten meaning differences. As for the genre-specific semantic fields, the altered position (towards or away from the prototypical center) or sheer absence of certain lexemes in the genre-specific fields seems to be an indicator for the general degree of formality as well as of the topic variety typical of the text type under study (e.g., Figure 3).

This paper thus not only contributes to the current state of the art in corpus-based translation studies by focusing on the semantic relationships between translations and original texts, but also methodologically by designing a new method for more statistically-based and semantics-oriented research in the field of corpus-based translation studies.

References

Aijmer, K., & Simon-Vandenbergen, A.-M. (2004). A model and a methodology for the study of pragmatic markers: the semantic field of expectation. *Journal of Pragmatics, 36*(10), 1781-1806.

Dagan, I., Itai, A., & Schwall, U. (1991). *Two languages are more informative than one.* Paper presented at the Proceedings of the 29th annual meeting of the Association for Computational Linguistics Berkeley, California.

Dyvik, H. (1998). A translational basis for semantics. In S. Johansson & S. Oksefjell (Eds.), *Corpora and cross-linguistic research: theory, method, and case studies* (pp. 51-86). Amsterdam: Rodopi.

Dyvik, H. (2004). Translations as semantic mirrors from parallel corpus to wordnet. In K. Aijmer & B. Altenberg (Eds.), *Advances in Corpus Linguistics* (pp. 311-326). Amsterdam & New York: Rodopi.

Greenacre, M. (2007). *Correspondence analysis in practice, Second edition.* . Boca Raton: Chapman & Hall/CRC.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data*. Dordrecht: Kluwer Academic Publishers.

Lefever, E. (2012). *ParaSense: parallel corpora for word sense disambiguation.* Ghent University, Ghent.

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

Simon-Vandenbergen, A.-M. (2013). English adverbs of essence and their equivalents in Dutch and French. *Advances in Corpus-Based Contrastive Linguistics: Studies in Honour of Stig Johansson, 54*, 83.

Vandevoorde, L., De Sutter, G., & Plevoets, K. (in press). On semantic differences between translated and non-translated Dutch. Using bidirectional parallel corpus data for measuring and visualizing distances between lexemes in the semantic field of inceptiveness. In J. Meng (Ed.), *Corpus-Based Study of Translation Lexis*. Sheffield: Equinox.

Exploring language specificity as a variable in interpreting performance: A corpus-based investigation of attributive modifying structures in Chinese-English consecutive interpreting

Binhua Wang
Hong Kong Polytechnic University
wangbinhua@hotmail.com

For a long period in its history, interpreting studies has focused on the exploration of cognitive processing in interpreting behaviours. However, an adequate description of interpreting behaviours and activities, as the disciplinary foundation of interpreting studies, requires not only the exploration of internal cognitive processing factors but also the examination of external social and cultural factors. The major shaping forces of interpreting products include: a) the interpreter's interpreting competence, b) cognitive condition on the site and c) norms of interpreting (Wang, 2012). For interpreting activities between languages involving wide differences in linguistic structure and cultural conceptualization, esp. between European and non-European languages, language specificity is also a major factor shaping interpreting products.

Interpreting into B, esp. between Chinese and a European language poses special challenges to interpreters, most of which can be attributed to language specificity. As Setton (1999: 53) states, among the ten most popular languages used for simultaneous interpreting, four pose special difficulties as source languages: "three (Chinese, English and Japanese) have a significant amount of left-branching or SOV structure, while Arabic exhibits Verb-first forms".

Although a few studies have touched upon the issue of language specificity in interpreting, previous attempts have focused solely on Japanese-English interpreting, while the treatment of language-specificity in Chinese as a major non-European language used widely in interpreting has seen virtually no systematic exploration.

Based on the Corpus of Chinese-English Interpreting for Premier Press Conferences (CEIPPC), the present paper is a descriptive study of the issue of language-specificity, esp. syntactic differences between Chinese and English discourses in interpreting and the relevant interpreting strategies employed by interpreters. The annotated corpus consists of 14 press conferences interpreted by seven professional interpreters in the consecutive mode, which are of much homogeneity in both forms and topics. The investigation is focused on how attributive modifying structures in Chinese is transformed to English in consecutive interpreting. Special attention is paid to the interpreters' handling of the influence of linguistic differences that are of much specificity to the Chinese-English language pair.

This study may shed new light on the role of language specificity as a factor shaping interpreting product and implies the necessity of considering it as a variable in the explanatory account of interpreting behaviours, esp. those between languages involving wide differences in linguistic structure and cultural conceptualization, esp. between European and non-European languages.

References

Setton, R. 1999. Simultaneous Interpretation: A Cognitive-pragmatic analysis. Amsterdam/Philadelphia: John Benjamins

Wang, B. 2012. A Descriptive Study of Norms in Interpreting – Based on the Chinese-English Consecutive Interpreting Corpus of Chinese Premier Press Conferences. *Meta*. 57 (1)

Wang, E. 2008. Interpreting into B: A comparative survey of three East-Asian countries (In Chinese). *Chinese Translators' Journal*. 2008 (1)

Investigating translator's notes: A corpus-based study

Ting-Hui Wen
Changhua University of Education, Taiwan
tinghuiwen@cc.ncue.edu.tw

The current study tries to investigate the phenomenon of copious translator's notes in Chinese translated texts and the strategies adopted by translators to add notes.

In Chinese translation, adding notes is quite common. Translator's notes can be inserted to explain certain people, places, history, social phenomena, allusions and puns. Sometimes translators can even add their own interpretations in the target texts. Translator's notes usually come in the forms of footnotes, endnotes or between parentheses in the texts

Translator's notes indicate the presence of translators. Venuti disagrees with the invisibility of translators, and he claimed that invisibility is translators' "self-annihilation," and translation could therefore be marginalized (1995:8). Hermans also stated that translation has a second voice, which is the translator's voice, and the translator's note is the most overt way to present translator's voice (1996: 27). Chao studied footnotes from a sociological perspective and treated notes as the practices of "thick translation" (2011: 17). Her corpus included Chinese translations of Angela Carter's novels published in Taiwan, and she categorized the notes into three categories: linguistics issues; intertextual features; and socio-cultural background details. In Lai's proposal of new translation of classic literature, she also emphasizes the presence of translators, and proposes that the translator's voice should be heard (2012: 3-10).

The current research investigates translator's notes using the Parallel Corpus of Chinese Mystery Fiction (PCCM). The PCCM is an extended corpus of the Comparable Corpus of Chinese Mystery Fiction (CCCM), which included translated and non-translated texts published in Taiwan from the year 2000 to 2005. The source texts are included in the original CCCM to enable further studies on different translation features. Only one translated text in the PCCM does not include any notes in any forms; seven out of the eight titles of translated mystery fiction included in the PCCM have translator's notes: four have footnotes, and three have notes between parentheses in the main texts.

The current study further investigate the different types of translator's notes to understand the strategies adopted by translators regarding what, when and why they add notes in their translations. Moreover, the different strategies adopted by different translators and publishers will also be investigated.

References

Chao, J. (2011) *Translational Footnotes and the Positioning of Unfamiliar Literature: Capital flow of translations of Angela Carter's novels in Taiwan*, Unpublished PhD thesis, The University of Manchester.

Hermans, T. (1996) "The translator's voice in translated narrative", *Target* 8 (1): 23-48.

Lai, S. T. (2012) "Translator as commentator: On the Translator's Notes by Woo Kuang Kien", *Compilation and Translation Review* 5 (2): 1-29.

Venuti (1995) The Translator's Invisibility: A History of Translation, New York and London: Routledge.

A trilingual parallel corpus-based contrastive study of the past tense in Spanish, English and Chinese

Meng-Hsin Yeh
NCKU, Taiwan
k26024055@ncku.edu.tw

Hui-Chuan Lu
NCKU, Taiwan
huichuanlu1@gmail.com

An-Chung Cheng
University of Toledo, USA
accheng99@gmail.com

This paper focuses on introducing the creation of a trilingual parallel corpus, CPEIC, by a research team at the Cheng Kung University in Taiwan and its application on a contrastive analysis of the past tense in Spanish, English and Chinese.

1 The construction and application of a parallel corpus

Among different types of corpora, the construction of parallel corpora benefits research in contrastive analysis, translation and language acquisition (e.g., Baker, 1993; Malmkjaer, 2005; Rabadán, Labrador & Ramón, 2009; Dimitrova, Koseska-Toszewa, Roszko & Roszko, 2010). Conducting contrastive analysis through parallel corpus also facilitates, particularly, the comparison and contrast among semantically similar however syntactically different phrases or sentences in two languages. Among the 32 exiting parallel corpora in the field[22], 12.5% (4/32) of them are related to English-Spanish and 16% (5/32) are related to English-Chinese. Nevertheless, there is no parallel corpus of Spanish-Chinese, nor a trilingual parallel corpus concerning the world's most spoken languages: English, Spanish and Chinese. Creating such a trilingual parallel will facilitate the research not only on contrastive linguistics, but also on second or foreign language acquisition.

This study examines a particular linguistic feature, the past tense. The past tense behaves differently in these three languages; there are two verb forms of past tense in Spanish, the preterite and imperfect; only one in English, past tense; and none in Chinese. Furthermore, the parallel corpus, CPEIC (*Corpus Paralelo de Español, Inglés y Chino* (Spanish)/Parallel Corpus of Spanish, English and Chinese) reflects the context of the Spanish language acquisition in Taiwan, in which Chinese is learners' native language, English is, typically, a second language (L2), and Spanish, the third language (L3). The creation of the trilingual parallel corpus and the findings of contrastive analysis through the CPEIC will provide useful implications for Spanish language teaching and learning. Thus, this paper will address the following two questions:

- What are the major features and functions of the CPEIC?
- How do three languages, Spanish, English and Chinese differ in the past tense in grammatical aspect, lexical aspect, and syntactic structure?

2 The creation of trilingual parallel corpus

In the process of constructing the CPEIC, collected data were imported into MySQL to be POS-tagged using TreeTagger for Spanish and English and CKIP for Chinese, and words were aligned through Giza++. A web-based user interface was designed by JavaScript along with JQuery, and the server side was programmed by PHP. The construction result[23] of the present stage includes Bible, fairy tales, and sources from the United Nations with both oral and written texts. These three sub-corpora contain 1,217,971 Spanish words, 1,190,081 English words and 1,543,580 Chinese words. The main features and functions are the compatibility across languages in Spanish, English and Chinese, word and sentence alignment, and POS-tagged information.

The search interface is divided into two sections, left and right. The conditions of search are set on the left hand side, including: (1) Sources of different sub-corpora, (2) three different languages, Spanish, English and Chinese (3) multiple keywords or conditions, (4) specific part of speech without specifying any keyword. On the right hand side appears the search result.

In the version of 2014, the following improvements have been made: (1) Increased speed of search and decreased shut-down frequency, (2) additional function of displaying POS-tagging, (3) improved function of displaying word alignment by highlighting words wherever the cursor is, and (4) enhanced compound queries.

3 Contrastive analysis of past tense in Spanish, English and Chinese

Given the situation in which words of Spanish, English and Chinese will be aligned in parallel within the same text in the CPEIC, one can easily compare and contrast the expressions of a same meaning. Similarities and differences in these three languages will be examined in terms of the following variables. First, grammatical aspects in three languages differ. Spanish has two expressive

[22] Lee, D. 2010. http://www.uow.edu.au/~dlee/CBLLinks.htm [2014-1-9]

23 http://140.116.245.228/TriApp/TriLin.html

ways, preterite and imperfect; English has one, simple past tense; and as for Chinese, there is no morphological aspect involved. However, Chinese has four aspectual markers, GUO, ZAI, ZHE and LE. Second, lexical aspects of verbs (state, activity, accomplishment and achievement) will be considered. Third, syntactic structure such as verb-object and temporal adverbs or phrases appear in the context will also be included in the discussion.

The paper will end with a discussion of cross-language influence in acquisition and implications of the CPEIC in teaching beginning and intermediate and advanced learners.

References

Baker, M. 1993. "Corpora in translation studies: An overview and some suggestions for future research". *Target* 7 (2): 223-243.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., and Roszko, R. 2010. "Application of multilingual corpus in contrastive studies (on the example of the Bulgarian-Polish-Lithuanian parallel corpus)". *Études Cognitives/Studia Kognitywne* 10: 217-240.

Malmkjaer, K. 2005. *Linguistics and the Language of Translation*. UK: Edinburgh University Press.

Rabadán, R., Labrador, B., and Ramón, N. 2009. "Corpus-based contrastive analysis and translation universals. A tool for translation quality assessment English Spanish". *Babel* 55 (4): 303-328.

A Corpus-based contrastive study of appraisal in English Military News

Zhang Ren
National University of Defence Technology
mollyzhang8025@hotmail.com

Ma Xiaolei
National University of Defence Technology
dennisma912@aliyun.com

1 Background

It is widely acknowledged that discourse has many different functions (Jakobson 1960). One important function of discourse is that it can be used to express feelings, attitudes and points of view. This function of discourse has been explored by many scholars using various terms (e.g. evaluation, appraisal, stance, etc.) and adopting different methodologies (e.g. Thompson and Hunston 2000; Martin and Rose 2003; Conrad and Biber 2000).

Among these approaches, appraisal studies are developing rapidly in recent years. Appraisal theory was originally put forward to develop ideas about the interpersonal metafunction in Systemic Functional Linguistics. According to Martin and Rose (2003: 22), "appraisal is concerned with evaluation: the kinds of attitudes that are negotiated in a text, the strength of the feelings involved and the ways in which values are sourced and readers aligned". Appraisal is envisaged as being composed of three subsystems: attitude, engagement and graduation. They are respectively concerned with what have traditionally been dealt with under the headings of "affect"; "evidentiality and epistemic modality"; "intensification and vague language" (Martin and White 2005: 2). Each subsystem consists of various subcategories. Take attitude for example, it can be further divided into "affect", "appreciation", and "judgment". And each of these can be further divided and thus a classification system of appraisal resources in discourse is suggested by appraisal theory.

2 Methodology

Appraisal in media discourse is a popular field of research in recent years (e.g. White 1998, 2006; Bednarek 2006), but these studies focus upon appraisal in news published in English speaking countries, and little attention has been paid to English news written by non-native speakers and published in countries where English is not the official language. This study sets out to examine and compare the appraisal resources employed in military news discourse posted on the website of American Department of Defence and the website of

Chinese Ministry of National Defence. The purpose is not only to identify the difference in the usage and patterning of appraisal resources between news discourse produced by native speakers and that produced by non-native speakers, but also to explore the different image building strategies of these two countries.

Two weeks of news in 2013 are respectively collected from the American Department of Defence website and the Chinese Ministry of National Defence website. These news reports mainly cover topics of institutional issues of the department, military operations, military exchanges, etc. A corpus consisting of two sub-corpora (63,641 and 61,071 tokens respectively) is built and appraisal resources (the graduation subsystem is not included in this study) are annotated manually by using labels in appraisal theory. WordSmith 5.0 is then applied to the quantitative analysis of the frequencies and distribution of various subcategories of appraisal resources in the two sub-corpora respectively. The two patterns of distribution are then compared and qualitative analysis is further conducted to look at more details and take the context into consideration.

3 Findings

The following characteristics of attitude can be found in the two sub-corpora. First, the general distribution pattern of the three subtypes of attitude (i.e. affect, appreciation and judgment) is the same across the two sub-corpora, with "appreciation" taking up the largest proportion and "affect" the smallest proportion. Second, there is a significant difference between the judgment values adopted in the two sub-corpora, with the positive judgment values outnumbering negative ones in the sub-corpus of American DoD news while the negative judgment values outnumbering positive ones in the sub-corpus of Chinese MoD news. Third, in both sub-corpora, the attitudinal values attributed to sources other than the author are adopted much more frequently than those attributed to the author.

Among the subcategories of engagement, the focus of study is put on "attribute" because of its frequent use in the two sub-corpora. It is found that the sub-corpus of American DoD news employs instances of "attribute" more frequently than the sub-corpus of Chinese MoD news. The sources quoted in the sub-corpus of American DoD news are largely American officers, while those quoted in the sub-corpus of Chinese MoD news range from Chinese officers to those of many other countries. The reporting verbs frequently used in both sub-corpora include *say*, *add*, *note* and *tell*. However, the sub-corpus of Chinese MoD news shows a preference for reporting verbs which are attitudinally loaded (e.g. *accuse*, *urge*, *hope*), while the sub-corpus of American DoD news tends to employ reporting verbs which are more neutral with regard to attitude (e.g. *explain*, *acknowledge*, *continue*).

A closer look at the two sub-corpora reveals that the American military reports are more subtle in its tone and more strategic in manipulating appraisal resources. It is found that the implicit realizations of attitude are often embedded with non-authorial inscriptions of attitude in the American military reports. This patterning helps to make the stances and opinions conveyed by the American military reports more difficult for the readers to detect and reject.

4 Conclusion

It can be concluded that although these two groups of English military news share a general patterning of appraisal resources, there are subtle differences between them which can be attributed to less skillfulness of Chinese reporters in applying English appraisal resources, as well as different strategies adopted to promote the image of military forces. It should be noted that the current study only focuses on military news and need to be extended to news of other topics (political, entertainment, etc.) to see if there are any different findings.

References

Bednarek, M. 2006. *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. London/New York: Continuum.

Conrad, S. & D. Biber. 2000. "Adverbial marking of stance in speech and writing". In Hunston, S. & G. Thompson (eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse* (pp. 57-73). Oxford: Oxford University.

Hunston, S. & G. Thompson. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University.

Jakobson, R. 1960. "Closing statement: Linguistics and poetics". In T. A. Sebeok (ed.), *Style in Language* (pp. 350-377). Cambridge: Cambridge University Press.

Martin, J. R. & Rose, D. 2003. *Working with Discourse*. London & New York: Continuum.

Martin, J. R. and White, P. R. R. 2005. *The Language of Evaluation: Appraisal in English*. London/New York: Palgrave/Macmillan.

White, P. R. R. 1998. "Telling Media Tales: the News Story as Rhetoric". PhD. Thesis. University of Sydney.

White, P. R. R. 2006. "Evaluative Semantics and Ideological Positioning in Journalistic Discourse". In Lassen, I. (ed.), *Image and Ideology in the Mass Media* (pp. 45-73). Amsterdam/Philadelphia: John Benjamins.

Translation and language change based on diachronic multiple corpora: a case study of modifiers in early modern Chinese

Zhao Qiurong
University of Science and Technology Beijing
qiurong_zhao@126.com

Wang Kefei
Beijing Foreign Studies University
kfwang126@126.com

Corpus-based translation studies have rendered great insights into translational phenomena in recent years such as translation-induced language contact and language change. Most of them are within the closely related language pairs, English-German (House 2003, 2006; Baumgarten et al. 2004, 2008; Steiner 2008; Kranich et al. 2011, 2012), English-Danish (Gottlieb 2004), English-Italian (Laviosa 2010), English-French (McLaughlin 2011) and so on, while little has been undertaken in the distant language pairs, such as the influence of English translation on Chinese text production. Modern Chinese has undergone dramatic changes during the New Culture Movement of the early 20th century at that time a flurry of translation activities known as Europeanization is hoped to have shouldered the mission of promoting modern Chinese and adopting new ideas and even new modes of thinking. During this period, modern Chinese gradually replaced classical Chinese and finally settled down to its present form. Many studies have focused on the Europeanized structures in Chinese, while most of them are relatively subjective based on typical but few examples and most of them are synchronic, without giving the necessary background about the time and social contexts of the translated Chinese.

Modifier is one of the most typical syntactic features of translated Chinese and one of the indicators influenced most by English-Chinese translation. Based on the diachronic multiple corpora, namely, a combination of comparable corpora (1910-1949), parallel corpora (1930-1949) and reference corpus (before 1905), the present paper attempts to explore the influence of translation on the change of modifiers, in the form of "*one+modifier+head noun"* (for instance, 一个*曾经帮助过我*的人（literary translation: one has once helped me subordinating maker person; *a man who had once helped me*). Modern Chinese has changed dramatically within a short time, multiple corpora with shorter time periods between them will give a more reliable account of change. By diachronic analysis, the study examines comparable literary works of approximately 2,100,000 words, covering a 40-year period from 1910 to 1949. The 40 years are divided into four sub-periods, 1915-1920, 1925-1930, 1935-1940 and 1945-1949, so as to reveal the delicate changes in modern Chinese and discover the nature of language change in different social contexts.

The examination will help to reveal the development of modern Chinese in terms of longer and more complicated modifier in the structure of "*one/the/this +modifier + head noun*" over time, in particular, what role translation has played during the developmental process.

The research questions are:

(1) What are the differences between translated Chinese and non-translated Chinese over time in terms of the length and structure of modifiers?

(2) How and to what degree, translation has influenced native Chinese text production in different social contexts?

The findings show that,

(1) The length of modifiers in translated Chinese in four periods is 5.85, 7.56, 7.85 and 7.75 respectively; the length in non-translated Chinese is 6.58, 6.97, 7.28 and 7.29 respectively. Generally speaking, the length of modifiers in both translated Chinese and non-translated Chinese is on the increase. In particular, the length of modifiers in translated Chinese is longer than that of non-translated Chinese in the latter three periods, and only the first period is with an exception. In the first period, Europeanization was actively advocated, and this period began to focus on "faithful translation", while "free translation" still had a big proportion. Furthermore, although most of the translators have accepted the idea of Europeanization, they are deeply influenced by *classical Chinese form* for their educational background, so it is hard for them to completely get rid of the trace of *the classical form* within a short time.

(2) The normal state of modifier in classical Chinese is short, but the findings show that the length of modifier in the reference corpus is 6.78, which is longer than that in the authentic Chinese in the first period. After careful investigation, it shows that the long modifier is separated by some punctuation, which is different from the translated version. For instance,

一个 方巾裲衫、十字披红、 金花插帽、

（one/ square scarf goffer,/ drape a band of red silk over his shoulders, / golden flowers on the hat,/

满脸酸文、 一嘴尖团字儿的 一个人（*from the reference corpus*）

be overfastidious in wording,/ a sharp-tongued character/ one person ）

A man with a square scarf goffer, a shawl of red silk draping over his shoulders and golden flowers stuck on his hat, which is accustomed to using archaism and often has a sharp tongue. (Translated by the present author)

(3) Close examination reveals that the authentic Chinese imitated the long modifier from the English subordinate clauses. Chinese is left-branching structure, while English is right-branching structure. In the period of advocating copying, the English subordinate clauses are translated into the premodifiers of Chinese, thus resulting in the modifier to become longer and more complicated.

(4) Translation is the catalyst and gateway in the development of modern Chinese, but how and to what degree it may function depends on many parameters, among which social and cultural contexts are important.

References

Baumgarten, N., J. House and J. Probst. 2004. "English as Lingua Franca in Covert Translation Processes".*The Translator*, 10 (1).83-108.

Baumgarten, N. and D. Özçetin. 2008. "Linguistic Variation Through Language Contact in Translation". In P. Kintana and N. Siemund, (eds.), *Language Contact and Contact Languages*. Amsterdam: John Benjamins Publishing Company. 293-316.

Gottlieb, H. 2004. "Anglicisms and Translation". In G. Anderman and M. Rogers (eds.), *In and out of English: For Better, for Worse?* Clevedon: Multilingual Matters LTD. 161-184.

House, J. 2003. "English as Lingua Franca and Its Influence on Discourse Norms in Other Languages". In G. Anderman and M. Roger (eds.), *Translation Today:Trends and Perspectives*. Clevedon: Multilingual Matters Ltd. 168-180.

House, J. 2006. "Covert Translation, Language Contact, Variation and Change." *SYNAPS*, 19. 25-47.

Kranich, S., V Becher,. and S. Höder. 2011. "A Tentative Typology of Translation-induced Language Change". In S. Kranich, V. Becher, S. Höder and J. House, (eds.), *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*. Amsterdam: John Benjamins Publishing Company, 11-43.

Kranich, S., J. House, and V. Becher. 2012. "Changing Conventions in English-German Translations of Popular Scientific Texts". In K. Braunmüller and C. Gabriel (eds.), *Multilingual Individuals and Multilingual Societies*. Amsterdam: John Benjamins Publishing Company. 315-334.

Laviosa, S. 2010. "Corpus-Based Translation Studies: 15 Years On". *SYNAPS*, 24.3-12.

McLaughlin, M. 2011. *Syntactic Borrowing in Contemporary French: A Linguistic Analysis of News Translation*. Oxford: Legenda.

Steiner, E. 2008. "Empirical Studies of Translations as a Mode of Language Contact". In Siemund, P. and N. Kintana, (eds.). *Language Contact and Contact Languages*. Amsterdam: John Benjamins Publishing Company. 317-345.