International Symposium

Parallel Corpora: Creation and Applications

PaCor 2021

Vitoria-Gasteiz, 23, 24 and 25 June 2021

Sponsors:

English and German Philology, Translation and Interpreting Department

Universidad del País Vasco | Euskal Herriko Unibertsitatea

LETREN FAKULTATEA
FACULTAD DE LETRAS

Faculty of Arts

Campus of Álava

Contents

About

The research group TRALIMA/ITZULIK GIC IT 1209-19 of the University of the Basque Country/Euskal Herriko Unibertsitatea (UPV/EHU) is pleased to organise the III International Symposium on Parallel Corpora, PaCor 2021. This conference will be carried out ONLINE on 23, 24, and 25 June 2021.

International Symposium PaCor 2021

PaCor 2021 aims to contribute to the scientific dissemination initiated by the research group SpatiAlEs, from the University of Santiago de Compostela, in 2016, later reinforced by the Instituto Universitario de Lenguas Modernas y Traductores (ULMYT), at the University Complutense of Madrid, in 2018. TRALIMA/ITZULIK together with the abovementioned research groups, and many others, take part in CORPUSNET. The goal of this network is the development of (parallel or comparable) corpus-based tools, applications and resources to satisfy needs in the realms of research, teaching and/or intercultural communication (http://corpusnet.unileon.es/).

Parallel corpora creation and exploitation are possible thanks to the collaboration of linguists, (computational) engineers, statisticians and a variety of language users (researchers, learners, translators, among others). While the latter report their needs for language use, as well as problems or challenges in cross-cultural communication, the former describe languages, at different levels, to observe what should be done to meet each ultimate purpose, thus defining a possible solution. It is the engineers and statisticians who give it shape by developing tools whose usefulness and usability be guaranteed. This interdisciplinary collaboration is as necessary as complex and fraught with challenges. This is indeed the case given the increasing range of language applications, not only among various languages but also in a variety of domains.

Originally, the aim of PaCor is twofold: 1) to identify challenges, from a variety of perspectives including contrastive linguistics and translation, to name a few, with the intention of extending applications to solve them; and 2) to provide a platform for presentation of projects on parallel corpora where Spanish is the pivot language. We hope this third edition adds on to the knowledge gained in previous editions, not only by giving awaiting answers but also by raising new questions that, altogether, enhance corpus linguistics in general and parallel corpora in particular. To this end, we would like PaCor 2021 to pay special attention to parallel corpora that feature, at least, one minority language.

Organising committee

Marlén Izquierdo (chair)	Zuriñe Sanz (secretary)
Olaia Andaluz	Elizabete Manterola
Raquel Merino	María Pérez
Alejandro Ros	Hugo Sanjurjo-González
Ana Tamayo	Naroa Zubillaga

Timetable

Wednesday, 23 of June

9:00–9:30		Opening (https://labur.eus/ICoQu)	
		Raquel Merino: Principal Investigator of Research Group TRALIMA/ITZULIK	**Cande Cabanillas: Vice-dean for Internships, Multilingualism and Assistanships (Faculty of Arts) & member of TRALIMA/ITZULIK**
10:00–11:00	KL	**Signe Oksefjell-Ebelling** University of Oslo (https://labur.eus/ICoQu)	Bidirectional parallel corpora: Challenges and possibilities
11:00–11:30		**Break**	
Chair		**Signe Oksefjell-Ebeling** (https://labur.eus/ICoQu)	**Marlén Izquierdo** (https://labur.eus/Z1PeL)
11:30–12:00	CT	**Josep Marco** Light verb constructions as a testing ground for the Gravitational Pull Hypothesis: An analysis based on COVALT corpus	**Noelia Ramón** Exploring near-synonyms through translation corpora: A case study on *begin* and *start* in the English-Spanish Parallel Corpus PACTRES
12:00–12:30	CT	**Rosa Rabadán** Light Verb constructions in translation. What corpora tell us	**Teresa Ortego** Las herramientas semiautomáticas de redacción y traducción español-inglés basadas en corpus: el ejemplo de GEFEM
12:30–13:00	CT	**Hui Chuan-Lu & An Chung Cheng** Study of Verb Tenses and Temporal Adverbs among Spanish, English, and Chinese	**Belén Labrador** Run away! The tip and the iceberg of core vocabulary
13:00–14:00		**Lunch**	
Chair		**Zuriñe Sanz** (https://labur.eus/ICoQu)	**Marlén Izquierdo** (https://labur.eus/Z1PeL)
14:00–14:30	CT	**João Almeida** Building a parallel corpus for Portuguese Tetum	**Timea Kovács** Simplification in English-Hungarian translated and interpreted texts in an inter-modal English-Hungarian "sub-corpus"
14:30–15:00	CT	**Li Biwei** Using multilingual parallel corpus for Journalistic Translation Research: (re)constructing national images via global news in English, Chinese and Spanish	**Raquel Lázaro** Application of Corpus Pattern Analysis for the study of face-threatening acts (FTAs) in telephone interactions mediated by an interpreter

Thursday, 24 of June

Chair		**Elizabete Manterola** (https://labur.eus/ICoQu)	**Naroa Zubillaga** (https://labur.eus/Z1PeL)
09:30–10:00	CT	**Camino Gutiérrez** English-Spanish dubbese vs. natural pre-fabricated orality: a corpus-based study of conversational markers	**Teresa Molés-Cases** –Es el fin –balbució–, „Das ist das Ende“, stammelte er. Reporting direct speech in Spanish and German
10:00–10:30	CT	**Aitziber Elejalde** Ciencia ficción: neología y (re)traducción. Un estudio de corpus transmedia	**Leonor Pérez** Looking at Flemish tapestries from the wrong side: Translation (Spanish-English) of cultural references in rural tourism hospitality industry web pages
10:30–11:00	CT	**Irene Hermosa** Opera audio description: A lexico-grammatical corpus analysis of Catalan and Spanish scripts	**María Teresa Sánchez** Expresión de la perspectiva del recipiente en la traducción alemán-español. Un estudio de las pasivas de *bekommen*, de *erhalten* y de *kriegen* en los datos del corpus PaGeS
11:00–11:30	**Break**		
11:30–12:30	KL	**Xavier Gómez Guinovart** Universidade de Vigo (https://labur.eus/ICoQu)	Las redes semánticas en la construcción y explotación de corpus paralelos[1]
12:30–14:00	**Lunch**		
Chair		**Hugo Sanjurjo** (https://labur.eus/ICoQu)	**Xavier Gómez** (https://labur.eus/Z1PeL)
14:00–14:30	CT	**Flavia De Camillis & Giovanni Contarino** Adapting machine translation for under-resourced languages: a first attempt for institutional German in South Tyrol	**Tian Mi & Rodrigo Muñoz** The Chinese-Spanish corpus of Journey to the West
14:30–15:00	CT	**Olga Bonetskaya; Dmitry Dolgov; Maria Frolova; Anastasia Politova; Anna Prykova** On word alignment of Russian-Chinese parallel corpora	**Anja Weingart; Georg A. Kaiser; Svenja Schmid** Creating a multilingual parallel corpora: the UV2 web application
15:00–15:30	CT	**Antonina Bondarenko** Verbless Questions in Multidirectional Parallel Corpora: A Contrastive Study	**Daniel Rojas** Elaboración de un corpus paralelo sobre movimiento causado en las lenguas romances

[1]This plenary lecture will be delivered in Spanish. To listen to its interpreting into English, open the interpreter's session as well: https://labur.eus/SIY3p

Friday, 25 of June

9:00–10:00	KL	**Nora Aranberri** University of the Basque Country (https://labur.eus/ICoQu)	Corpus paraleloak eta itzulpen automatikoa: aukerak, erronkak eta… euskara[2]
10:00–10:30	**Break**		
Chair		**Marlén Izquierdo** **(https://labur.eus/ICoQu)**	**Nora Aranberri** **(https://labur.eus/Z1PeL)**
10:30–11:00	CT	**Andrea Götz** Native speakers use more connectives? A corpus-based examination of L1 and L2 Hungarian to English interpreting	**Gert de Sutter** Text and translation in context: embracing qualitative data in corpus-based translation studies
11:00–11:30	CT	**Naroa Zubillaga** Enriching the MUST project: Basque and EN, DE or ES translation pairs	**Zuriñe Sanz** Compilation of DIY parallel corpora in the translation classroom using TAligner
11:30–12:00	**Break**		
Chair		**Hugo Sanjurjo, Marlén Izquierdo, Zuriñe Sanz** **(https://labur.eus/ICoQu)**	
12:00-13:30	WS	**Johannes Graën** Universität Zürich	Parallel corpora: tools and applications
13:30–14:00	**Break**		
14:00-15:00	WS	**Johannes Graën** Universität Zürich	Parallel corpora: tools and applications
15:00	**Closing (https://labur.eus/ICoQu)**		

KL: Keynote Speaker, CT: Contributed Talk, WS: Workshop

[2] This plenary lecture will be delivered in Basque. To listen to its interpreting into English, open the interpreter's session as well: https://labur.eus/SIY3p

List of Abstracts

Keynote Speakers

Parallel corpora in machine translation: opportunities, challenges, and... Basque

Dr. Nora Aranberri KL

University of the Basque Country (UPV/EHU)

Parallel corpora are vital to the development and evaluation of many natural language processing applications. In many cases, however, compiling suitable parallel resources poses an enormous challenge. In this talk, we will focus on machine translation (MT) and consider a number of situations at different stages of the development and implementation cycle where parallel corpora play a key role. We will first concentrate on the development stage, and specifically consider the features of the data required to build the systems. We will look into ways in which researchers have tried to generate the parallel corpora, discussing examples of targeted manual generation and automatic generation, including the implications of back-translation. Secondly, we will examine the requirements of the parallel corpora used in the implementation stage to help users take full advantage of MT and also corpora compiled to draw conclusions on MT use by professional translators and regular users. Throughout the talk we will present specific examples where Basque is involved, allowing us to highlight the implications of working with a low-resource minority language.

Semantic networks in the construction and exploitation of parallel corpora

Dr. Xavier Gómez Guinovart KL

Universidade de Vigo

In this talk, I will explain the research on parallel corpora recently conducted in the Seminars of Computational Linguistics at the University of Vigo. I will focus on the use of lexico-semantic information, as provided by WordNet, in the construction and exploitation of the CLUVI and SensoGal corpora respectively. This combination of resources is possible in both directions, namely, from the parallel corpus to WordNet and from WordNet to the parallel corpus.

On the one hand, it is possible to apply on the parallel corpora a variety of equivalents extraction techniques that widen the lexical coverage of the wordnets of the languages under alignment. It is also possible to benefit from parallel corpora to obtain contexts of use, for WordNet, of the concepts compiled in the net, provided that the corpus is previously processed with a suitable semantic framework.

On the other, WordNet may be used in the alignment of parallel corpora at the lexical level as well as in their lexico-semantic annotation. For example, the graph technique of semantic relations in WordNet is used for constricting semantic taggers able to disambiguate, lexically, parallel corpora. Another resource used for this purpose has been English language corpus SemCor, semantically annotated by the team who developed English WordNet in Princeton.

I will attempt to provide a wide overview of the many facets of the research in progress for the audience to perceive the benefits of lexico-semantic annotation in the construction and exploitation of parallel corpora.

References

Gómez Guinovart, X. Solla Portela, M.A. (2020). Construction of a WordNet-based multilingual lexical ontology for Galician. In M. J. Domínguez Vázquez, M. Mirazo Balsa C. Valcárcel Riveiro (Eds.) *Studies on Multilingual Lexicography*, pp. 179-196. De Gruyter, Berlin and Boston.

Gómez Guinovart, X. (2019). Enriching parallel corpora with multimedia and lexical semantics: From the CLUVI Corpus to WordNet and SemCor. In I. Doval M. Teresa Sánchez Nieto (Eds.), *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*, pp. 141-158. John Benjamins, Amsterdam.

Simões, A. Gómez Guinovart, X. (2018). Extending the Galician wordnet using a multilingual Bible through lexical alignment and semantic annotation. In P. Rangel Henriques, J. P. Leal, A. Menezes Leitão X. Gómez Guinovart (Eds.) *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*, pp. 14:1-14:13. Schloss Dagstuhl/Leibniz-Zentrum fuer Informatik, Dagstuhl.

Gómez Guinovart, X. Solla Portela M.A. (2018). Building the Galician wordnet: Methods and applications. *Language Resources and Evaluation*, 52 (1) 317-339.

Bidirectional parallel corpora: Challenges and possibilities

Dr. Signe Oksefjell Ebeling KL

University of Oslo

In this talk I will start by outlining some of the main challenges relating to the use of bidirectional parallel corpora for contrastive research, offering some insights from my own experience of compiling and using parallel corpora of this kind. These challenges notwithstanding, I will then move on to describe the potential of bidirectional parallel corpora and give a snapshot of some of the possibilities they offer. More specifically, I will give examples of different kinds of contrastive studies that have benefitted from the bidirectional corpus design devised by Stig Johansson (Johansson Hofland 1994). The selection of studies discussed, mainly from my own research, will range from lexical and lexico-grammatical studies of predefined items and patterns in two languages to more exploratory studies of n-grams.

Contributed talks

Building a parallel corpus for Portuguese Tetum

J. João Almeida and Alberto Simões

CT

University of Minho

Building parallel corpus in general is a challenging task. It gets much harder when we deal in minority languages, where de amount of texts and language resources is small and the amount of translations is not so big.
In this article we discuss strategies used do build to parallel corpus: Portuguese-Tetum.
The set of tools build and adapted to deal with Tetum, and to clean the set of resources found, will also be discussed.
Finally we will present a set of valuable auxiliar language resources created and used in the precess.

Introduction

When the amount of bitexts is small, the traditional techniques often fail and we have the look for more specific strategies:

- What are the Bitext that we know that exist? (Example: The East Timor Constitution, and some of the juridical codes were built with the support of the Portuguese Government)
- What are the bilingual websites for TP-PT (example "http://timor-leste.gov.tl/").
- What are the best translators for the language? (Example: Esperança)
- What are the dictionaries / glossaries that build a bridge between the source / target languages?
- What are the universal projects that both languages share? (example UDHR)
- What are the common subtitles?
- What are the didactic texts for the language?
- Can we build bitexts using triangulation aproaches? (transitive bitexts)
- Can we find comparable translations?
- text containing both languages (bilingual texts, translations or comparable).

In the main article we will discuss case studies covering all the different situations:

Extracting and aligning Bilingual WebPages Example: East Timor Secretaria de Estado de Arte e Cultura (http://www.cultura.gov.tl/)

Texts with multi-language versions Example: Convention for the safe guarding of the intangible Cultural Heritage (Jornal da República) (http://www.cultura.gov.tl/sites/default/files/serie_i_no_19_unesco_3.pdf)

Complex Bilingual and Multilingual Texts Example: "O Anjo de Timor", Sophia de Mello Breyner Using glossaries and dictionaries Example: Glossary of the Land Law Program (including terms and definitions in for languages Portuguese, Tetum, English and Indonesian)

Extracting from existing corpora Example: Tetum from Linguee and Glosbe

Specific tools

The following specific tools, created to help in minority languages resource creation, will be discussed:

- Fine grain language Identification.
- Language annotation and splitting.
- Corpora Work-Flow.

Discussion of results

The current state of the Tetum-Portuguese parallel corpus Will be presented and the future steps discussed.

References

José João Almeida, Sílvia Araújo, Nuno Carvalho, Idalete Dias, Ana Oliveira, André Santos, and Alberto Simões. The Per-Fide corpus: A new resource for corpus-based terminology, contrastive linguistics and translation studies. In Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira, editors, *Working with Portuguese Corpora*, pages 177–200. Bloomsbury Publishing, April 2014.

José João Almeida and Alberto Simões. Automatic parallel corpora and bilingual terminology extraction from parallel websites. In Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff, editors, *BUCC2010 – 3rd Workshop on Building and Using Comparable Corpora*, lrec2010, pages 50–55, Valletta, Malta, May 2010.

Rui Brito and José João Almeida. A workflow description language to orchestrate multi-lingual resources. In *3rd Symposium on Languages, Applications and Technologies, SLATE 2014, June 19-20, 2014 - Bragança, Portugal*, volume 38 of OASICS, pages 77–83. Schloss Dagstuhl - Leibniz- Zentrum fuer Informatik, 2014.

Rui Brito, José João Almeida, and Alberto Simões. Processing annotated TMX parallel corpora. In *IberSpeech 2014 — VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop*, pages

188–197, Las Palmas de Gran Canaria, Spain, November 2014.

Nuno Ramos Carvalho, Jose Joao Almeida, Maria João Varanda Pereira, and Pedro Rangel Henriques. Probabilistic synset based concept location. In SLATe'12 — *Symposium on Languages, Applications and Technologies*, volume 21, pages 239–253. OASIC – Open Access Series in Informatics, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, June 2012.

Pedro Carvalho and José João Almeida. Mlt-prealigner: a tool for multilingual text alignment. In *3rd Symposium on Languages, Applications and Technologies, SLATE 2014, June 19-20, 2014 - Bragança, Portugal*, volume 38 of OASICS, pages 283–290. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2014.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Alberto Simões, José João Almeida, and Nuno Ramos Carvalho. Defining a probabilistic translation dictionaries algebra. In Luís Correia, Luís Paulo Reis, José Cascalho, Luís Gomes, Hélia Guerra, and Pedro Cardoso, editors, *XVI Portuguese Conference on Artificial Inteligence - EPIA*, pages 444–455, Angra do Heroismo, Azores, September 2013.

Alberto Simões, José João Almeida, and Nuno Ramos Carvalho. Defining a probabilistic translation dictionaries algebra. In *XVI Portuguese Conference on Artificial Inteligence - EPIA*, pages 444–455, Angra do Heroismo, Azores, September 2013.

Alberto Simões, Xavier Gómez Guinovart, and J. João Almeida. Enriching a Portuguese WordNet using synonyms from a monolingual dictionary. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia, may 2016. European Language Resources Association (ELRA).

Alberto Simões, B. Sacanene, Alvaro Iriarte, J.J. Almeida, and J. Macedo. Towards a morphological analyzer for the umbundu language. volume 83, 2020.

Alberto M. Simões and J. João Almeida. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September 2003.

Verbless Questions in Multidirectional Parallel Corpora: A Contrastive Study

Antonina Bondarenko

CT

Université de Paris

What does contrastive analysis reveal about the typological differences in the use and meaning of the absence of the verb in English and Russian? The paper combines contrastive and corpus methods in order to explore the semantic and pragmatic characteristics of verbless questions in English and Russian reciprocal and third-language parallel corpora.

The difficulty of detecting absence automatically has generally limited studies of verbless sentences from taking a corpus-linguistic approach. Grammatical structures are ideally a matter for parsed corpora (Gilquin, 2002), but the retrieval of structures without a verbal predicate is typically hampered by verb-centric syntactic modeling and fixed annotation, as shown by (Landolfi et al., 2010) in a survey of several existing parsed corpora. Tagged corpora do not include a zero marker (Loock, 2016: 33). Moreover, little attention has been paid to the accuracy of automatic retrieval of the structures in existing studies, which as shown in (Bondarenko, 2019) is a serious issue for the verbless phenomenon. As a result, studies of the verbless phenomenon tend to focus on particular predefined structures that are more easily searchable in existing corpora or turn to examples that are invented and fragmented from various sources.

We use a new method designed for accurate verbless sentence retrieval (Bondarenko, 2019). The reciprocal parallel corpus consists of:

- L1 Russian L2 English: Fyodor Dostoyevsky's Russian novel (1880) and two English translations – The Brothers Karamazov (Pevear Volokhonsky, 1990) and The Karamazov Brothers (Avsay, 1994),

- L1 English L2 Russian: Harold Pinter's English play The Caretaker (1960) and its Russian translation – (Doroshevich, 2006)

- L3 Russian L3 English: two Russian and two English translations of Albert Camus' French novel L'Etranger (1946) – (Adamovich, 1966), (Gal, 1968), The Stranger (Gilbert, 1946), The Stranger (Ward, 1988).

The raw texts were sentence segmented, morpho-syntactically tagged and aligned at the sentence level across multiple translations with the help of Trameur (Fleury Zimina, 2014). The latter processes custom segmented data, permits automatic correction of a large portion of tagging errors, classifies the custom segments according to the presence of the verb, permits the visualization of verbless sentences aligned with multiple translations in their original context, as well as statistical analysis against a reference corpus of verbal sentences. Following extraction, verbless questions and their translation correspondences were manually annotated for: antecedent-based verbal ellipsis, verbal translation correspondence, discourse type, direct and indirect speech act and question type in accordance with (Celle et al., 2019; Celle, 2018). Verbal translation correspondences of verbless sentences are analyzed in terms of potential implicature (Bondarenko Celle, 2020).

The corpus is analyzed from three perspectives. First, from a monolingual perspective, translations are

treated as genuine samples of language in their own right, following (Zanettin, 2014; Olohan, 2002; Baker, 1996; Biber, 1993), and compared with the translations and the originals in the corresponding language in terms of key word analysis and manually annotated categories. Secondly, combining the contrastive analysis principles of (Guillemin-Flescher, 2003) with criteria for reliable parallel corpora (Nádvorníková, 2017; Stolz, 2007; McEnery Xiao, 2008; Malmkjaer, 1998), we look for reciprocal correspondence patterns that reoccur across multiple translations, texts and directions. Verbal correspondences of non-antecedent based verbless questions are correlated with speech acts and question types. Thirdly, following (Zanettin, 2013; Baker, 1993), a third-language sub-corpus consisting of Russian and English translations from French is added in attempt to control for source language interference on the data.

The objects under study are questions defined on a semantic level, as opposed to syntactic interrogative clauses (Huddleston, 1994). The defining syntactic feature of English interrogatives, the subject-auxiliary inversion, is inapplicable to verbless structures. In Russian, intonation is the principle way of distinguishing questions (Comrie, 1984). Profound typological differences, including the fact that Russian is known for permitting the most liberal use of verbless sentences among the Indo-European family, and English for its dependence on the verb phrase, make comparison of the two languages particularly relevant for the study of verbless phenomenon (Stassen, 2013; Kopotev, 2007; Leech, 2004)

Previous analyses justify the focus on the non-elliptical verbless question, i.e. not based on a recoverable syntactic antecedent (Bondarenko Celle, 2019; Bondarenko, 2019; Bondarenko, 2018). As shown, differences in the frequency of the verbless phenomenon cannot be explained by the productivity of syntactic ellipsis: despite Russian formally allowing more productive verbal ellipses thanks to its morphological case system, a higher frequency of antecedent-based ellipsis is found in English. Non-elliptical verbless utterances dominate in both languages. Furthermore, the verbless phenomenon is statistically linked to direct speech, discourse genre and lexical markers of common ground, which emphasizes their importance in interactive settings.

Analysis of translation correspondences shows that questions, compared to other sentence types, are particularly sensitive to the verb. Regardless of language type (source, translation or third-language translation), verbs vary more frequently in correspondences of verbless questions than in all of the other sentence types combined.

In both languages, non-elliptical verbless questions are found to correlate with indirect speech acts, i.e. questions where the speaker is not requesting information from the addressee, such as the rhetorical question in (1).

(1) {Speaker interrupts interlocutor in a debate. Excommunication is used as an argument.}

Как это отлучение, что за отлучение?

kak èto otlučenie, čto za otlučenie

what this excommunication, what particle excommunication

What do you mean excommunication? What excommunication?

Rather than syntax, our results suggest that contributing to the explanation of verbless sentence frequency differences between the languages is a typological difference in the pragmatic use of the verb in questions. English verbal correspondences of verbless Russian questions are found to correlate with direct speech acts across all language types. This suggests that in English the verb allows to distinguish between questions as direct versus indirect speech acts, whereas in Russian this distinction is not related to the verb. We present evidence that the absence of the verb may be used as a grammatical marker in English to signal that a questioning speech act is indirect.

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology* (pp. 233–250). John Benjamins.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP & Translation*. John Benjamins.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.

Bondarenko, A. (2018). Predication Transformation: A parallel corpus-based study of Russian verbless sentences and their English translations. *Annales Universitatis Mariae Curie- Sklodowska Philologiae: La Linguistique Contrastive*, 36(1), 43–54.

Bondarenko, A. (2019). A corpus-based contrastive study of verbless sentences: Quantitative and qualitative perspectives. *Studia Neophilologica*, 1–24. https://doi.org/10.1080/00393274.2019.1616221

Bondarenko, A., Celle, A. (2019). Traduire l'absence: Les questions à prédicat zéro dans un corpus parallèle russe et anglais. *Des Mots Aux Actes: Sémantique, Sémiotique et Traductologie*, 7, 341–364.

Bondarenko, A., Celle, A. (2020). Verbless sentences in L'Étranger: A French-Russian- English contrastive study. In E. Corre, D.-T. Do-Hurinville, & H.-L. Dao (Eds.), *The Expression of Tense, Aspect, Modality and Evidentiality in Albert Camus's "L'Étranger" and Its Translations* (Vol. 35, pp. 325–352). John Benjamins.

Celle, A. (2018). Questions as indirect speech acts in surprise contexts in English. In D. Ayoun, A. Celle, L. Lansari (Eds.), *Tense, Aspect, Modality, Evidentiality: Crosslinguistic Perspectives* (pp. 213–238). John Benjamins.

Celle, A., Jugnet, A., Lansari, L., Peterson, T. (2019). Interrogatives in surprise contexts in English. In N. Depraz A. Celle (Eds.), Surprise at the Intersection of Phenomenology and Linguistics. John Benjamins.

Comrie, B. (1984). Interrogativity in Russian. In W. Chisholm, L. Milic, J. Greppin (Eds.), *Interrogativity: A Colloquium on the Grammar, Typology and Pragmatics of Questions in Seven Diverse Languages* (pp. 7–46). John Benjamins.

Fleury, S., Zimina, M. (2014). Trameur: A framework for annotated text corpora exploration. In L.

Tounsi R. Rak (Eds.), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (pp. 57–61). Dublin City University and Association for Computational Linguistics. http://www.aclweb.org/anthology/C14-2013.pdf

Gilquin, G. (2002). Automatic retrieval of syntactic structures: The quest for the Holy Grail. *International Journal of Corpus Linguistics*, 7(2), 183–214.

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

Huddleston, R. (1994). The Contrast between Interrogatives and Questions. *Journal of Linguistics*, 30(2), 411–439.

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

Landolfi, A., Sammarco, C., Voghera, M. (2010). Verbless clauses in Italian, Spanish and English: A Treebank annotation. In S. Bolasco, I. Chiari, L. Guiliano (Eds.), *JADT 2010: Proceedings of the 10th International Conference on Statistical Analysis of Textual Data* (pp. 1187–1194). LED Edizioni Universitarie.

Leech, G. N. (2004). *Meaning and the English Verb* (3rd ed.). Longman.

Loock, R. (2016). *La traductologie de corpus*. Presses Universitaires du Septentrion.

Malmkjaer, K. (1998). Lover thy Neighbour: Will Parallel Corpora Endear Linguists to Translators? *Meta*, 43(4), 534–541. https://doi.org/10.7202/003545ar

McEnery, A., Xiao, R. (2008). Parallel and comparable corpora: What is Happening? In G. Anderman M. Rogers (Eds.), *Incorporating Corpora: The Linguist and the Translator* (pp. 18–31). Multilingual Matters.

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

Olohan, M. (2002). Corpus linguistics and translation studies: Interaction and reaction. *Linguistica Antverpiensia*, 1, 419–429.

Stassen, L. (2013). Zero copula for predicate nominals. In M. S. Dryer M. Haspelmath (Eds.), *The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology*. http://wals.info/chapter/120

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

Zanettin, F. (2013). Corpus methods for descriptive translation studies. *Procedia - Social and Behavioral Sciences*, 95, 20–32.

Zanettin, F. (2014). Corpora in translation. In J. House (Ed.), *Translation: A Multidisciplinary Approach* (pp. 178–199). Palgrave Macmillan.

Keywords: verbless questions, indirect speech acts, pragmatics, contrastive analysis, parallel corpora, automatic retrieval of absence, English, Russian

On word alignment of Russian-Chinese parallel corpora

Olga BONETSKAYA; Dmitry DOLGOV; Maria FROLOVA; Anastasia POLITOVA; Anna PYRKOVA CT

HSE University, Computer Science Faculty; Independent AI researcher; Independent researcher; School of Liberal Arts, Nanjing University; HSE University, School of Asian Studies

Word alignment of parallel corpora is defined as finding word-to-word relationships between bitexts already aligned on a sentence level. It is a fundamental task in both natural language processing and linguistics. In the former, it is essential for a range of downstream tasks like knowledge transfer from high-resource to low-resource languages, namely projecting input formatting and annotations, and also for evaluation of machine translation quality. In the latter, it is of great help to the learners of foreign languages. Beginners can search for clear examples of a single word or grammar usage. Translators may check translations of the problematic idioms or collocations and choose a better solution.

The current study aims at developing a word-aligned Russian-Chinese dataset, evaluating and comparing various machine learning algorithms that can produce word-to-word alignment automatically, and developing a web tool to identify word/expression translations in context within a Russian-Chinese parallel corpus - RuZhCorp (Durneva et al., 2020). RuZhCorp was created in 2016 by a team of sinologists and computational linguists as a branch of the Russian National Corpus. Having got its independent version at the Higher School of Economics (Moscow) website in 2019, so far, RuZhCorp has accumulated 1 070 texts and 3.5 million words both in Russian and Chinese. Text types include fiction, news, and official documents, but as currently, 80% of the corpus is fiction, the present study focuses primarily on the fiction genre.

Building a dataset, we initially developed an alignment manifesto, a set of high-level principles stipulating that alignment should be based on word representation in the discussed languages and not on the context. In other words, only tokens with clear semantic correspondence in both languages should be aligned, while those added or omitted due to context necessity, literal purposes or to prevent tautology in a language should not. The manifesto itself has been based on existing works and approaches in the field of word alignment. We consulted the experience of manual word alignments over six language pairs (combinations between Portuguese, English, French and Spanish) described in the paper “Building a golden collection of parallel Multi-Language Word Alignments“ (Graça et al., 2008). The outcome comprised one hundred examples for each language pair extracted from Europarl Corpus. Subsequently, the researchers have compiled a detailed manual alignment guideline that assumes division into sure and possible alignments, and this experience was adopted by our team. Additionally, our manual alignment approach includes some unique features. We use Google Sheets instead of the Annotation tool software designated in (Graça et al., 2008). Then, similarly to the predecessors, we defined a set of specific rules to achieve a better representation of searched words in the bilingual corpora. The classifiers, which are widely used in the Chinese language, were one of the stumbling blocks in the alignment process; therefore, we have set that only those that have a clear match in the corresponding language are to be linked. Similar rules were developed for prepositions, auxiliary verbs, modal particles, etc. Therefore, we both proposed a theoretical roadmap for the unification of the word alignment principles in other language pairs and elaborated the precise alignment rules for the Russian-Chinese language pair.

We have created a framework for manual annotation of the parallel corpora with word alignments

that allows capturing many-to-many relationships in a way that is easily understandable both by humans and machines. This framework distinguishes between sure and possible alignment links and incorporates an iterative peer-review process to guarantee the uniform quality of the resulting dataset. The framework was applied to 125 randomly selected sentence pairs that were used for the evaluation of artificial intelligence algorithms. By analogy with the previous research, we used variable conditional designations for alignment points such as “1” (i.e. sure alignment point for existing machine alignment), “2” (i.e. possible alignment point for existing machine alignment), “n” and “p” (for similar cases when there is a correspondence between the word pair that has been omitted by machine alignment), and “q” (i.e. questionable alignment point that is to be discussed) performed by the first annotator. In case there was the machine markup for the non-corresponding word pairs, a first annotator should put the “d” letter meaning deletion of an inappropriate alignment point. A second annotator performs the second iteration and puts “11” and “22” for sure and possible alignment points relatively and deletes “d”s if he or she agrees with the first annotator, otherwise “q” should be put. Such a two-step iteration process allows for a relatively high rate of consistency.

As for the programming part of the alignment, historically automatic word alignment was mostly done using statistical methods. The expectation-maximization algorithm was first proposed by Dempster et al. (1977) and implemented for word alignment under the name of IBM models by Brown et al. (1993). Och and Ney (2003) created a tool called GIZA++ that remains a common benchmark till now. Later, several deep learning approaches had been proposed. Yang et al. (2013) used a DNN (deep neural network) to discriminatively learn bilingual word embeddings. Bahdanau, Cho and Bengio (2015) used a DNN to jointly learn to align and translate. Stengel-Eskin et al. (2019) used supervised learning to extract alignments from the attention module of a Transformer DNN.

Using our novel dataset, we have selected, reparameterized, applied, and compared several machine alignment approaches. The main method was extracting alignment from contextualized word embeddings (Dou and Neubig, 2021) of deep learning language models like BERT and LaBSE with further fine-tuning on the parallel corpus. BERT (Bidirectional Encoder Representations from Transformers) is a language model pre-trained on unlabeled monolingual texts in different languages (Devlin et al., 2018). LaBSE (Language-agnostic BERT Sentence Embedding) is a model that combines masked language model and translation language model that is trained on parallel data (Feng et al., 2020). EM (expectation-maximization) was used as a statistical benchmark that neural models were evaluated.

In our pilot study, we have evaluated word alignments that can be extracted directly from publicly available versions of BERT and LaBSE. Then, we additionally trained those models on the Russian-Chinese parallel corpus. As a final step of the research, we plan to fine-tune those models on additional gold standard data (500 sentence pairs more) marked manually by human annotators. To compare the quality of the models, we use the AER (Alignment Error Rate) metric introduced by Och and Ney (2000). All RuZhCorp texts (3.5 million words) are used for training. And when training on annotated sentences from the novel dataset, AER is calculated on the rest of the dataset.

As of now, our project team has compiled a fully developed parallel corpus of 125 sentence pairs along with over 10 well-established and proven rules for Russian-Chinese and Chinese-Russian word alignment. When comparing algorithms on this dataset prior to fine-tuning, LaBSE achieves the best AER of 32-36% (depending on whether only sure or both sure and possible links are considered), and BERT follows with AER of 38-39%. Due to the lack of previous works on Russian-Chinese word alignment, we have compared our results with the results for other comparable pairs of non-similar languages that include one European and one East-Asian language. Li et al. (2019) list 36.57% as their

best result for Chinese-English; Dou and Neibig (2021) show AER of 37.4% for Japanese-English while providing a much lower AER of 13.9% for Chinese-English. The data shows that we are in line with previous research on similar language pairs; in fact, our results may become a valuable benchmark for future research on Russian-Chinese word alignment. Such results already allow for the development of the above-mentioned tool for language scholars. Further, we plan to improve them using two levels of fine-tuning (on parallel corpora and additional gold standard data).

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *arXiv*:1409.0473.

Peter F. Brown, Stephen A. Della-Pietra, Vincent J. Della-Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation. *Computational Linguistics*, 19(2):263-311.

Arthur P. Dempster , Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1-38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*:1810.04805.

Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. *arXiv*:2101.08231

Sofia. P. Durneva, Yulia N. Kuznetsova, and Kirill I. Semenov. 2020. Russian-Chinese Parallel Corpus of RNC: Problems and Perspectives. *Proceedings of the 10th International Conference "Russia and China: history and perspectives for cooperation"*, 633–640. (София П. Дурнева, Юлия Н. Кузнецова, Кирилл И. Семенов. 2020. Русско-китайский параллельный корпус НКРЯ: проблемы и перспективы. X Международная научно-практическая конференция «Россия и Китай: история и перспективы сотрудничества», 633-640.)

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv*:2007.01852.

João Graça, Joana Paulo Pardal, Luisa Coheur, and Diamantino Caseiro. 2008. Building a golden collection of parallel Multi-Language Word Alignments. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng Shuming Shi. 2019. On the word alignment from neural machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1293–1303.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 440–447.

Franz Josef Och and Hermann Ney. *2003. A systematic comparison of various statistical alignment models.* Computational Linguistics, 29(1):19–51.

Elias Stengel-Eskin, Tzu-Ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 910–920.

Nan Yang, Shujie Liu, Mu Li, Ming Zhou, and Nenghai Yu. 2013. Word alignment modeling with context dependent deep neural network. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 1), 166-175.

Adapting machine translation for under-resourced languages: a first attempt for institutional German in South Tyrol

Flavia De Camillis; Antonio Giovanni Contarino

Eurac Research; University of Bologna

Keywords: under-resourced domain; machine translation; institutional translation; parallel corpus; legal terminology

The setting of our study is South Tyrol, a small province in Italy, where Italian and German are co-official languages. Despite German being a well-resourced, pluricentric language, the South Tyrolean variety is to a certain extent under-resourced. This is because institutional German in South Tyrol is strongly bound to the Italian legal system, which makes it substantially different from other German varieties as concerns legal and administrative terminology. Standardised and recommended terminology for South Tyrol is collected in the freely accessible informative system *bistro*[1] (Ralli Andreatta, 2018). Public institutions publish many of their documents (e.g. laws, invitations to tender, resolutions) in both Italian and German systematically since the 1990s. However, they are one of the few sources of legal and institutional discourse in South Tyrolean German, making it a narrow domain. For machine translation research, this is a well-known limitation (Koehn Knowles, 2017; Michon et al., 2020; Skadiņa et al., 2010). It is also the primary reason why South Tyrolean civil servants cannot rely on mainstream MT-tools for their translations. Their output may be quite good standard German thanks to the progress made by neural network models (Barrault et al., 2020; Vaswani et al., 2017), but legal and administrative terms are frequently mistranslated (De Camillis, 2021; Wiesmann, 2019).

Against this background, our exploratory study is the first attempt – to the best of our knowledge – of tailoring an MT-system to South Tyrolean institutions. From our experiments, we expect improvements particularly in what concerns legal and administrative terminology. For this pilot phase, we chose an adaptive NMT-system, ModernMT, as its adaptation approach allows on-the-fly fine-tuning of a pre-trained baseline model based on an in-domain adaptation set (Bertoldi, Caroselli, et al., 2018). To set the tests, we collected existent legal and administrative resources in South Tyrolean German, consisting of published documents and local terminology. The same resources are also available in Italian. Firstly, we created a parallel corpus of local legislation scraping the public database LexBrowser[2]. 4987 texts were collected, aligned and accurately cleaned up and filtered using deterministic rules in order to retain high-quality sentence pairs. Cleaning operations consisted in removing segment-internal noise and correcting hyphenated words, whereas filtering operations included discarding bad sentence pairs according to several noise classes (wrong language, sentence length ratio, duplicates, etc.). At the same time, we collected and cleaned TMs from the central translation bureau of the local administration. Overall, we totalled approximately 243k translation units (11.6m tokens), as well as around 10k terms in German from the system bistro. Finally, we fed the ModernMT system with this data as TM files.

For our experiments, we first used a test set consisting of 2k segments from the LexBrowser corpus. The results reveal considerable improvements both in terms of automated metrics, achieving 50.61 BLEU (+20.30 BLEU over the ModernMT baseline system), and with regard to the translation of legal and administrative terminology. Further analyses in relation to legal terminology correctness and

[1] http://bistro.eurac.edu/
[2] http://lexbrowser.provinz.bz.it/

adequacy are still ongoing. With our exploratory study, we hope to pave the way for an in-depth research, aiming at creating an MT-system for the translating institutions of South Tyrol (Koskinen, 2008).

References

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., … Zampieri, M. (2020). Findings of the 2020 Conference on Machine Translation (WMT20). *Proceedings of the 5th Conference on Machine Translation (WMT)*, 1–55.

Bertoldi, N., Caroselli, D., & Federico, M. (2018). The ModernMT Project. In *Proceedings of the21st Annual Conference of the European Association for Machine Translation 28–30 May 2018 Universitat d'Alacant*, Spain (p. 345). https://rua.ua.es/dspace/bitstream/10045/76096/1/EAMT2018-Proceedings_48.pdf

De Camillis, F. (2021). *La traduzione non professionale nelle istituzioni pubbliche dei territori di lingua minoritaria: Il caso di studio dell'amministrazione della Provincia autonoma di Bolzano* [Doctoral thesis]. Università di Bologna.

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28–39. https://doi.org/10.18653/v1/W17-3204

Koskinen, K. (2008). *Translating Institutions. An Ethnographic Study of EU Translation*. St. Jerome.

Michon, E., Crego, J., & Senellart, J. (2020). Integrating Domain Terminology into Neural Machine Translation. *Proceedings of the 28th International Conference on Computational Linguistics*, 3925–3937. https://doi.org/10.18653/v1/2020.coling-main.348

Ralli, N., & Andreatta, N. (2018). Bistro – ein Tool für mehrsprachige Rechtsterminologie. trans-kom, 11(1), 7–44. Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., & Tufiş, D. (2010). Analysis and evaluation of comparable corpora for under resourced areas of machine translation. *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, LREC 2010, 6–14. https://doi.org/10.13140/2.1.2852.3520

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wiesmann, E. (2019). Machine translation in the field of law: A study of the translation of Italian legal texts into German. *Comparative Legilinguistics*, 37, 117–153. https://doi.org/10.14746/cl.2019.37.4

Text and translation in context: embracing qualitative data in corpus-based translation studies

Gert De Sutter

CT

Ghent University

Since the early 1990s, the availability of translational corpora has extensively contributed to the identification of the specific linguistic features of translated text in comparison to their source texts and comparable non-translated texts (see De Sutter Lefer 2019 for a critical overview). Nevertheless, recent empirical translation studies (e.g., Halverson 2015, De Sutter et al. 2012, Lefer 2020) agree that most traditional corpora are no longer suited to account for a thorough understanding of the (cognitive and social) mechanisms that shape the language used in translated texts. In order to further uncover the sociocognitive circumstances under which texts and translation are produced, compilers of (parallel) corpora are encouraged to develop new-generation corpora which are "more carefully designed to take consideration of translators' backgrounds and the circumstances of text production" (Kotze 2020: 356).

The present research project responds to this invitation for more qualitative data in corpus-based translation studies by introducing the Dutch Parallel Corpus 2.0 (DPC 2.0): a bidirectional parallel corpus of expert translations for Dutch><English and Dutch><French language pairs. The corpus, which readopts the main compilation and design principles of its predecessor (Macken, De Clercq Paulussen 2011), at the time of writing contains 2.75 million words and is furthermore sentence-aligned, lemmatized and POS-tagged by means of the state-of-the-art natural language processing toolkit Stanza. DPC 2.0 distinguishes itself from traditional parallel corpora through its considerable amount of metadata about the translators (e.g., gender, education, experience) and the translation projects (e.g., L1/L2 translation, software used, degree and type of revision), next to the traditional metadata about the texts and translations themselves (e.g., source and target language, intended audience, intended goal, register).

DPC 2.0 allows researchers from various disciplines to adopt a fine-grained approach to linguistic research on translations and their source texts in which the underlying, extra-linguistic context plays an important role. The output of each search query can in fact be filtered according to a large variety of text-related, translation-related and translator-related criteria, as well as a flexible combination of multiple criteria. As a result, end-users of DPC 2.0 are enabled to carry out descriptive-comparative analyses of, for instance, varying translator profiles or translational contexts. In this talk, we will present the results of a first general exploration of the corpus in terms of general statistics (a.o. frequency of different POS tags, TTR, lexical density…), taking into account the main categories represented in the metadata.

References

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

De Sutter, G., & Lefer, M.-A. (2019). On the need for a new research agenda for corpus-based translation studies : a multi-methodological, multifactorial and interdisciplinary approach. *Perspectives-studies in translation theory and practice*, 28(1), 1–23.

Halverson, S. L. (2015). Cognitive Translation Studies and the Merging of Empirical Paradigms. The Case of 'literal Translation.' *Translation Spaces*, 4(2), 310–40.

Kotze, H. (2020). Converging what and how to find out why: An outlook on empirical translation studies. In L. Vandevoorde, J. Daems & B. Defranq (Eds.), *New Empirical Perspectives on Translation and Interpreting* (pp. 333-371). Routledge.

Lefer, M.-A. (2020). Parallel corpora. In M. Paquot S. Th. Gries (Eds), *A Practical Handbook of Corpus Linguistics* (pp.257–282). Springer.

Macken, L., De Clercq, O., Paulussen, H. (2011). Dutch parallel corpus: a balanced copyright-cleared parallel corpus. *META*, 56(2), 374–390.

Ciencia ficción, neología y (re)traducción. Un estudio de corpus transmedia

Aitziber Elejalde Sáenz CT

Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)

La ciencia ficción es un género en auge en la actualidad, tanto en el panorama literario como en el audiovisual. Las nuevas editoriales o sellos especializados y las plataformas de visión bajo demanda están apostando cada vez más por este tipo de contenido. El éxito que las novelas de este género han tenido entre los lectores ha fomentado su adaptación a la pequeña pantalla.

Aunque sea prácticamente imposible encontrar una definición de ciencia ficción con la que todo el mundo esté de acuerdo, lo cierto es que una de sus características más destacadas es la abundancia de neologismos (Stockwell, 2000), ya sean términos creados expresamente para una obra en concreto, o palabras ya establecidas, pero de uso delimitado a este género.

El presente estudio forma parte de la investigación que estoy llevando a cabo para mi tesis doctoral. El objetivo es analizar la traducción al español peninsular de los neologismos de una novela de ciencia ficción, Altered Carbon. Esta novela, publicada en inglés en 2001 por el escritor británico Richard Morgan y encuadrable dentro del subgénero del ciberpunk, ganó el premio Philip K. Dick en 2003. En español se publicó una primera traducción, *Carbono alterado*, en 2005 y una retraducción, Carbono modificado, en 2016. En 2018 Netflix estrenó la serie basada en esta primera novela bajo el título *Altered Carbon*.

Nos encontramos, pues, ante un caso de retraducción activa (Pym, 1998) y transmedia, en la que uno de los principales intereses será ver las diferencias entre las técnicas empleadas por los traductores literarios, así como entre las obras literarias y la audiovisual.

Para ello, siguiendo la propuesta de Juan Jesús Zaro (2007), se compilará un corpus con la novela original en inglés, las dos traducciones editadas en España, así como el guion y los subtítulos originales de la obra audiovisual y sus respectivas traducciones para doblaje y subtitulado. La herramienta empleada para compilar este corpus será TAligner 3.0, un programa escrito en Java que permite limpiar los textos rápidamente y etiquetarlos a nivel de párrafo y oración, así como alinear múltiples textos (Sanz, 2018).

Este primer análisis descriptivo permitirá determinar los diferentes enfoques traductológicos empleados a la hora de abordar la traducción de los neologismos. Además, se seleccionarán diferentes ejemplos para la realización de un estudio de recepción posterior cuyo objetivo es evaluar el nivel de comprensión y aceptabilidad de la audiencia.

Referencias bibliográficas

Kalogridis, L. (Productora ejecutiva). (2018-2020). *Altered Carbon* [Serie de televisión]. Skydance Television; Mythology Entertainment.

Morgan, R. (2001). *Altered Carbon*. Gollancz.

Morgan, R. (2005). *Carbono alterado*. (Trad. M. Tombetta y E. Gutiérrez). Minotauro.

Morgan, R. (2016). *Carbono modificado*. (Trad. J. Barranquero). Gigamesh.

Pym, A. (1998). *Method in Translation History*. St. Jerome Press.

Sanz Villar, Z. (2018). Diseño, descripción y análisis de un corpus multilingüe (alemán-español-euskera). *TRANS. Revista de Traductología*, 22, 133-148.

Stockwell, Peter (2000). *The Poetics of Science Fiction*. Routledge.

Zaro Vera, J. J. y Ruiz Noguera, F. (2007). *Retraducir. Una nueva mirada. La retraducción de textos literarios y audiovisuales*. Miguel Gómez Ediciones.

Native speakers use more connectives? A corpus-based examination of L1 and L2 Hungarian to English interpreting

Andrea Götz CT

Károli Gáspár University

Keywords: corpus-based interpreting studies, connectives, retour interpreting, relay interpreting, direct interpreting, EP interpreting, L2 discourse

Ideally, interpreters work into their mother tongue. However, retour interpreting has been the norm for small languages (Gentile Albl-Mikasa 2017), and is becoming a market reality (de la Iglesia and Opdenhoff 2014; Donovan 2010). Despite its ubiquity, retour interpreting is still under-researched, especially its implications for L2 discourse. This study addresses this gap by examining the use of connectives in a specially constructed parallel corpus composed of Hungarian to English L1 and L2 interpreting at the European Parliament (EP).

Although there is little research on connectives in retour interpreting, previous studies indicate certain trends. In L2 discourse, discourse markers have been shown to be less frequent (Götz 2013). On the other hand, connectives are generally frequent in interpreted discourse (e.g. Defrancq et al. 2015). Trainee interpreters working in retour have been observed to add connectives frequently in a bid to maintain cohesion (Gumul 2017), and can even be expressly instructed to do so (Wu and Liao 2018:203). Since L2 interpreting causes greater cognitive load (Chen 2020), it is plausible that maintaining cohesion would pose a challenge. Retour interpreting is also assumed to contain more interference (see Chmiel, Janikowski, and Cieślewicz 2020), which could also impact the use of connectives.

The parallel corpus of this study has been aligned by hand and compiled using Sketch Engine. It consists of (1) Hungarian to English direct interpreting (by L1 speakers), (2) retour interpreting (by L2 speakers), and (3) relay interpreting (by L1 speakers, mostly Hungarian to German to English). The corpus comprises 36,679 words and is c. 4 hours 41 minutes long. The retour interpreting section is made up of 46 speeches, and is approximately 1 hour 33 minutes long, containing 12,047 words. The study examines the frequency and the rate of addition of the following connectives: *as a result*, *but*, *however*, *nevertheless*, *now*, *so*, *that is*, *why*, *therefore*, *though*, *thus*, *well*, *yet*.

According to the results, while connectives are the most frequent in relay interpreting, retour interpreters use connectives more frequently than their English-speaking counterparts. This difference, however, is not due to a higher rate of addition but rather to retour interpreters transferring more items from source speeches than L1 interpreters. By percentage, retour interpreters use the most translated items, making the percentage of translated to added items almost equal (51.54% to 48.46%). By comparison, in both direct and relay L1 interpreting, 65% of items are added. Frequency also varies item to item. *Now*, for example, is significantly more frequent in L1 than in L2 interpreting, while the opposite is true for *therefore*. The first discrepancy is due to a lack of a direct Hungarian counterpart for the item *now*, the second, however, is caused by a lower rate of translation of the Hungarian item *ezért* ('for this reason') with *therefore*, despite *ezért* being more frequent in the source texts of L1 than of L2 interpreters. The findings of this study indicate that L1 and L2 interpreting differs in terms of connective frequency due to cross-linguistic differences, and highlight the need to incorporate L2 considerations into interpreter training.

References

Chen, Sijia. 2020. 'The Impact of Directionality on the Process and Product in Consecutive Interpreting Between Chinese and English: Evidence from Pen Recording and Eye Tracking'. *Journal of Specialised Translation*, no. 34: 100–117.

Chmiel, Agnieszka, Przemysław Janikowski, and Anna Cieślewicz. 2020. 'The Eye or the Ear?: Source Language Interference in Sight Translation and Simultaneous Interpreting'. *Interpreting. International Journal of Research and Practice in Interpreting* 22 (2): 187–210. https://doi.org/10.1075/intp.00043.chm.

Defrancq, Bart, Koen Plevoets, and Cédric Magnifico. 2015. 'Connective Items in Interpreting and Translation: Where Do They Come From?' In *Yearbook of Corpus Linguistics and Pragmatics* 2015, edited by Jesús Romero-Trillo, 3:195–222. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17948-3_9.

Gentile, Paola, and Michaela Albl-Mikasa. 2017. '"Everybody Speaks English Nowadays". Conference Interpreters' Perception of the Impact of English as a Lingua Franca on a Changing Profession'. *Cultus* 10: 53–66.

Götz, Sandra. 2013. Fluency in Native And Nonnative English Speech. *Studies in Corpus Linguistics*, volume 53. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Gumul, Ewa. 2017. 'Explicitation and Directionality in Simultaneous Interpreting'. *Linguistica Silesiana*, 311–29.

Iglesia, María Brander de la, and Jan-Hendrik Opdenhoff. 2014. 'Retour Interpreting Revisited: Tuning Competences in Interpreter Education'. *Current Trends in Translation Teaching and Learning* 1: 4–43.

Wu, Yinyin, and Posen Liao. 2018. 'Re-Conceptualising Interpreting Strategies for Teaching Interpretation into a B Language'. *The Interpreter and Translator Trainer* 12 (2): 188–206. https://doi.org/10.1080/1750399X.2018.1451952.

English-Spanish dubbese vs. natural pre-fabricated orality: a corpus-based study of conversational markers

Camino Gutiérrez Lanza CT

Universidad de León

Translation equivalence is traditionally identified with visual phonetics, lip-syncing (Fodor 1976), and isochrony (Whitman-Linsen 1992: 22) in dubbing (Gutiérrez Lanza 2000). However, the creation of pre-fabricated orality (fake spontaneous conversation among characters) includes additional language-related features such as discourse markers (Baños Piñero and Chaume 2009, Bruti 2019), which are often formally dissimilar across languages. The faulty use of various translation techniques (word-for-word translation, elision, etc.) very often results in the over/underuse of conversational markers. This is often done in favour of isochrony, but moves away from the recreation of natural pre-fabricated orality, negatively affecting acceptability, tenor and audience engagement (Rabadán and Gutiérrez Lanza 2020), creating unsuccessful communication patterns (Rabadán 2008 and 2010). However, over the years, Spanish audiences seem to have developed a high level of tolerance and *dubbese* (Corrius 2005; Romero Fresco 2006 and 2009; Pérez González 2007; Chaume 2007 and 2020) does not seem to disrupt their viewing experience.

This presentation reports on work-in-progress on one of the main problem-triggers in the recreation of natural pre-fabricated orality: conversational markers (*bueno*, *hombre*, *mira*, *oye*, *ya*, etc.) (Martín Zorraquino and Portolés Lázaro 1999). We aim a) to confirm whether they are overused or underused in English-Spanish dubbing and b) to suggest alternative solutions. To this end, we have started to compile a translated film script subcorpus as part of CETRI (Corpus del Español TRaducido del Inglés – Corpus of Spanish Translated from English) (ACTRES and TRACE 2019). CETRI contains materials from 2010 onwards distributed into two major subcorpora: fiction, roughly 19 million words, and non-fiction, with over 9 million words (as of December 2020). Texts have been PoS-tagged, and other annotation layers are being added gradually to allow for more refined searches. Translated scripts constitute a tiny part of the fiction subcorpus, roughly 50,000 words.

For this pilot study, we have focused on the film Jack Reacher (McQuarrie 2012). The "isochronised" TT2 (14,824 words), has been aligned with the intermediate translation, TT1 (12,799 words), and with the ST script (35,066 words), using TAligner 3.0 (Gutiérrez Lanza and Alonso 2011, MINECO 2019). In addition, CETRI materials have been compared with original Spanish data obtained from the equivalent subcorpus (*guiones*) by the Real Academia CORPES XXI (Corpus del Español del Siglo XXI – Corpus of 21st Century Spanish). Preliminary results indicate that most conversational markers tend to be used differently (they are either overused or underused) in English-Spanish translated scripts in relation to non-translated scripts in Spanish, which complies with isochrony restrictions, but often hampers the reproduction of realistic pre-fabricated orality in dubbing. They also suggest that dubbese can be avoided and pre-fabricated oraliy favoured if better options are available.

References

ACTRES and TRACE. 2019. Corpus del Español TRaducido del Inglés (CETRI). https://actres.unileon.es/internal/general_login/?url=/herramientas/cetri Accessed 28 Dec. 2020.

Baños Piñero, R. and Chaume, F. 2009. *Prefabricated orality. a challenge in audiovisual translation*.

InTRAlinea. Special issue: *The Translation of Dialects in Multimedia*. http://www.intralinea.org/specials/article/1714

Bruti, S. 2019. Spoken Discourse and Conversational Interaction in Audiovisual Translation. In Pérez González, L. (ed.). *The Routledge Handbook of Audiovisual Translation*. 192-208.

Chaume F. 2020. Dubbing. In Bogucki Ł. and Deckert M. (eds.). *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan, Cham. DOI: https://doi.org/10.1007/978-3-030-42105-2_6

Chaume, F. 2007. Quality Standards in Dubbing: a Proposal. *TradTerm* 13. 71-89.

Corrius, M. 2005. The Third Language: A Recurrent Textual Restriction that Translators Come across in Audiovisual Translation. *Cadernos de Traduçao*, 16, 147–160.

Fodor, I. 1976. *Film Dubbing*. Hamburg: Buske.

Gutiérrez Lanza, C. 2000. *Traducción y censura de textos cinematográficos en la España de Franco: doblaje y subtitulado inglés-español (1951-1975)*. León: Universidad de León.

Gutiérrez Lanza, C. Alonso, J. 2011. The TRACE Corpus Aligner: Developing a new electronic tool for language researchers. *III Congreso Internacional de Lingüística de Corpus. CILC 2011*. Universitat Politècnica de València. 7-9 April.

Martín Zorraquino, M.A. and Portolés Lázaro, J. 1999. Los marcadores del discurso. In Bosque, I. and Demonte, V. (eds.). *Gramática descriptiva de la lengua española*, vol. 3. Madrid: Espasa-Calpe.

McQuarrie, C. 2012. *Jack Reacher*. https://www.imdb.com/title/tt0790724/?ref_=nv_sr_srsg_0

Pérez González, L. 2007. Appraising Dubbed Conversation. Systemic Functional Insights into the Construal of Naturalness in Translated Film Dialogue. *The Translator* 13-1. 1-38.

Rabadán, R. and Gutiérrez Lanza, C. 2020. Developing Awareness of Interference Errors in Translation: An English-Spanish pilot study in popular science and audiovisual transcripts. *Lingue e Linguaggi* 40. 137-168.

Rabadán, R. 2008. Refining the Idea of 'Applied Extension.' In Pym, A.; Shlesinger, M. and Simeoni, D. (eds.). *Beyond Descriptive Translation Studies*. Amsterdam: John Benjamins. 103–117.

Rabadán, R. 2010. Applied Translation Studies. In Gambier, Y. and van Doorslaer, L. (eds.). *Handbook of Translation Studies*. Volume 1. Amsterdam: John Benjamins. 7–11. Real Academia de la Lengua Española. 2019. CORPES XXI. https://www.rae.es/recursos/banco-de-datos/corpes-xxiAccessed28Dec.2020.

Romero Fresco, P. 2006. The Spanish Dubbese: A case of (un)idiomatic Friends. *The Journal of Specialised Translation* 6. 134-151.

Romero Fresco, P. 2009. Naturalness in the Spanish Dubbing Language: A case of not-so-close Friends. *Meta*, 54/1. 49-72.

MINECO. 2019. CorpusNet. TAligner: http://corpusnet.unileon.es/herramientas-tecnicas Accessed 28 Dec. 2020.

Whitman-Linsen, C. 1992. *Through the Dubbing Glass: The synchronization of American motion pictures into German, French, and Spanish*, Peter Lang, Frankfurt am Main and New York.

Opera audio description: A lexico-grammatical corpus analysis of Catalan and Spanish scripts

Irene Hermosa Ramírez CT

Universitat Autònoma de Barcelona

Audio description (AD) is an audiovisual translation modality and media accessibility service which conveys the visual and otherwise inaccessible information for blind and visually-impaired audiences and users. Recent studies have taken an interest in empirically defining the specificities of AD from a corpus linguistics approach, most notably in the contexts of film (Reviers, 2017; Matamala, 2018) and museum AD (Jiménez Hurtado Soler Gallego, 2015; Perego, 2019). The present paper aims to expand this approach to opera AD.

Opera, a multimodal art form by nature, features some characteristics which set it apart from other audiovisual genres. Namely, its live nature, the concoction of musical, visual, dramatic and verbal codes, and the linguistic barrier which is otherwise overcome by surtitles. It is hypothesised that these idiosyncrasies will have an effect on the lexico-grammatical patterns of opera AD, distinguishing it from other AD modalities and general language samples in general. In order to test such claims, the author conducts a comparable corpus study of AD scripts delivered at the Liceu opera house in Barcelona (2007-2020) and at the Teatro Real in Madrid (2015-2019).

Design-wise, the sample gathers 13 scripts in Spanish from the Teatro Real and 15 scripts in Catalan from the Liceu opera house. For every script, a distinction has been established between the audio introduction (AI) and the AD "throughout" the performance itself. AIs are a "framework by which to understand the play" (Fryer Romero-Fresco, 2014, p. 9) and they are delivered before the start of the performance. The corpus is therefore divided into four subcorpora: AI in Catalan, AI in Spanish, AD in Catalan and AD in Spanish. The chosen software tools for the analysis of the corpus were Sketch Engine, WordSmith Tools 4.0, and JMP for the statistical analysis.

For the analysis, we define a number of criteria to assess the lexical (and some grammatical) features of opera AD. Among the most relevant results, all subcorpora are of a high lexical density. That is, a large proportion of the tokens in the corpus are lexical words. In terms of lexical variation, the standardised type-token ratio remains below 50%. As for legibility, average sentence length and word length are computed and the former yields an interesting outcome: there is a significant difference between all subcorpora except for the AI subcorpora ($t(51) = 1.20 = p = .23$). The paper goes on to explain such differences and commonalities with examples in context. Lastly, a semantic analysis is performed on the basis of generated frequency lists for open-class words. Namely, the most frequent nouns, verbs and adjectives are semantically tagged following the UCREL analysis system (Rayson *et al.* 2004). The semantic results, as well as the results of all measurements discussed above, are compared to previous corpus studies on other AD modalities.

References

Fryer, L., Romero Fresco, P. (2014). Audiointroductions. In A. Maszerowska, A. Matamala, P. Orero (Eds.), *Audio description: New perspectives illustrated* (pp. 11-28). Amsterdam: John Benjamins.

Jiménez Hurtado, C., Soler Gallego, S. (2015). Museum accessibility through translation: a corpus study

of pictorial AD. In J. Díaz Cintas, J. Neves (Eds.), *Audiovisual translation. Taking stock* (pp. 277-298). Newcastle upon Tyne: Cambridge Scholars Publishing.

Matamala, A. (2018). One short film, different audio descriptions. Analysing the language of audio descriptions created by students and professionals. *Onomázein*, 41, 185-207.

Perego, E. (2019). Into the language of museum audio descriptions: a corpus-based study. *Perspectives. Studies in Translatology*, 27(3), 333-349.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In L. Guthrie (Ed.), *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks* (pp. 7-12). Paris: European Language Resources Association.

Reviers, N. (2017). *Audio-description in Dutch: A corpus-based study into the linguistic features of a new, multimodal text type*. Antwerp: University of Antwerp. Unpublished PhD thesis.

A comparative analysis of English–Hungarian translated and interpreted texts in an inter–modal English–Hungarian "sub–corpus"

Timea Kovács CT

Károli Gáspár University of Reformed Church

Until now, relatively few inter–modal corpuses, including written and spoken texts and their interpreted and translated counterparts, have been devised. Nevertheless, the development of the EPTIC corpus was started by Bernardini et al. (2016) with a view to filling up this gap in the field of Corpus Linguistics. In the scope of their research, the authors built a multilingual (including English, Italian, French, Polish, and Slovenian languages) corpus including the original speeches of European Parliamentary sessions in 2011, their verbatim (written) reports, the source language translations made on the basis of the verbatim reports, as well as the simultaneously interpreted texts with English being in focus. The aim of the EPTIC project is to examine and compare lexical simplification, introduced by Laviosa (1998a, 1998b), in translated and interpreted texts in different language pairs and directions.

The aim of this paper is to illustrate how lexical simplification functions in the process of English–Hungarian translation and interpreting through the micro-analysis of aligned English–Hungarian translated and interpreted texts taken from the large EPTIC corpus.

In the analysis conducted in the English–Hungarian language EPTIC "sub-corpus", I am looking for answers to the question whether the text translated from English into Hungarian is lexically simpler than the interpreted one. On the basis of the results of the research hitherto conducted (Kovács, 2020), it can be stated that the sentences in the text interpreted from English into Hungarian are much shorter than in the translated text. As sentence length is a feature of lexical simplification, it can be concluded that the English–Hungarian interpreted text is lexically simpler than its translated counterpart. On the basis of the above results, I aim to examine through qualitative analysis what linguistic strategies adopted in simultaneous interpretation result, if they do, in the simplification (and in some cases, a loss) of content.

References

BERNARDINI, Silvia, FERRARESI, Adriano, MILICEVIC, Maja: From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective, *Target*, 28(1), 2016, 61–86.

KOVÁCS, Tímea: Korpusznyelvészet és tolmácsolás (Corpus Linguistics and Interpreting), in: *Korpusznyelvészet és nyelvi közvetítés (Corpus Linguistics and Interlinguistic Mediation)*, (eds. Mária Adorján and Tímea Kovács), L'Harmattan: Budapest, 2020, 73–91.

LAVIOSA, Sara (ed.): Special issue – The corpus-based approach: A new paradigm in translation studies, *Meta*, 43(4), 1998a.

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

Run away! The tip and the iceberg of core vocabulary

Belén Labrador CT

Universidad de León

Some words belonging to the core vocabulary of English are apparently easy to learn; for example, most learners of English, even beginners, would claim to know the word RUN, as they can only see the tip of the iceberg. However, knowing a word involves much more than roughly understanding its meaning and parallel corpora can be regarded as submarines that help us explore the bulk under the surface. In this translation-based study, motion expressions made up of the verb RUN and a satellite, "entity which acts as a spatial reference point for the motion/ location of the figure" (Talmy 2007:71), are examined under the microscope of P-ACTRES parallel corpus (Izquierdo et al 2008).

Drawing on a previous corpus-based paper on motion in English and Spanish (Author 2018), this piece of research takes RUN as the starting point, as it was found to be the most frequent verb in all the instances of crossed transposition. The study included all the verbs expressing manner of movement which collocated with the satellites selected. Crossed transposition (Molina and Hurtado Albir 2002) implies a double shift of part-of-speech from the source text to the target text and is the expected type of transfer between a satellite-framed language like English and a verb-framed language like Spanish, e.g. in *he rode away – se alejó al galope*, *rode* and *al galope* express manner; *away* and *alejarse*, path. Yet, crossed transposition ranked fourth in the list of translation solutions from English into Spanish. Leaving out either manner, e.g. *I climbed down – bajé* or path, e.g. *they jumped down – saltaron*, and expressing both in the verb, e.g. *which blew up – que explotó*, were the preferred options.

In this follow-up research, all the occurrences of the lexeme RUN, i.e. *run*, *runs*, *running* and *ran* have been analysed and their occurrences in expressions where RUN is followed by a satellite have been compared with their corresponding translations, in order to test whether crossed transposition ranks higher or lower than other translation solutions and to gain more general insight into the uses of this word. The analysis has applied Pym's typology (2018) and the results show that, although crossed transposition (or *perspective change* in Pym's terms) is frequent, RUN follows the general tendency in Spanish for *density change* - using only a verb either encompassing both meanings, path and manner, or making one of them implicit; in both cases the information is "spread over [. . .] less textual space" (Pym 2018: 57). After them comes *copying structure* - expressing manner through a verb and path through a satellite; finally, other minor translation solutions are *cultural correspondence* and *text tailoring*. The findings also include the most frequent translations for each combination of RUN plus satellite and cross-register differences between the fiction and the non-fiction subcorpora. Additionally, the most frequent collocates of the transitive uses of RUN have been identified. The final aim, with teaching purposes in mind, is to give a bigger picture of the iceberg underneath a high-frequency, basic little word like RUN.

Keywords: parallel corpus, translation solutions, motion, RUN, English-Spanish

References

Author. 2018.

Izquierdo, M., Hofland, K., and Reigem, Ø. 2008. The ACTRES Parallel Corpus: an English-Spanish Translation Corpus. *Corpora*, 3(1): 31–41.

Molina, L. and Hurtado Albir, A. 2002. Translation Techniques Revisited: A Dynamic and Funcionalistic Approach. *Meta*, 47(4): 498-512.

Pym, 2018. A Typology of Translation Solutions. *The Journal of Specialised Translation*, 30: 41-65.

Talmy, L. 2007. Lexical Typologies. In T. Shopen (ed.) *Language Typology and Syntactic Description*. Vol III: Grammatical Categories and the Lexicon (2nd edition), 66-168.

Application of Corpus Pattern Analysis for the study of face-threatening acts (FTAs) in telephone interactions mediated by an interpreter

Raquel Lázaro Gutiérrez CT

Universidad de Alcalá – Grupo FITISPos-UAH

Corpus linguistics, defined by McEnery and Wilson (2004:1) as the study of language based on examples of real-life language use, has been very productive in many of the branches of linguistics, including translation. The advantages that corpus methodology offers to interpreting studies are varied (Shlesinger, 1998), as it offers the possibility of obtaining information on grammar, lexical patterns, lexical density, discourse patterns, etc.

Regarding the analysis of FTAs with a corpus methodology, it is worth mentioning that the nature of the pragmatic aspects of language implies studying units that go beyond words or terms. For this reason, we opted for Corpus Pattern Analysis (CPA, Hanks, El Maarouf, Oakes, 2018) which not only focuses on Multi-Word Expressions (MWE), but also associates their meaning with the concrete use of the expressions, based on aspects such as phraseology or collocations. CPA is mainly based on the idea that, while words are associated with a multitude of meanings and are therefore ambiguous when decontextualised, most patterns have only one meaning. These patterns are associated with implicatures (de Schryver, 2010; Hanks, El Maarouf, Oakes, 2018). The meaning of the pattern is the primary implicature and, depending on the context and interpretation of the pattern, it may contain an indeterminate number of secondary implicatures.

While CPA has been used for lexical, semantic and syntactic analyses, it has not yet been applied to pragmatics. The project CM/JIN/2019-040 applies CPA to the analysis of FTAs in a multilingual corpus of telephone conversations between insurance operators and policyholders involving interpreters to find out:

- Which is the frequency of occurrence of the different types of FTAs?
- Who produces, receives and is affected by the FTAs?
- Which impact do FTAs have on the discourse and performance of interpreters?

In this contribution we aim to present the results of the validation of CPA as a useful methodology for our purposes. The corpus under analysis was compiled thanks to an agreement between the University of Alcalá and a Spanish telephone interpreting company. For quality purposes and with the consent of all parties, all conversations are recorded by the company. The sample for this study was randomly extracted from the bank of recordings of the company after an anonimization process and consists of 345 recordings of interactions with bilateral telephone interpreters in Spanish and Chinese, English, French, German, Italian and Russian, which took place from May 2017 to June 2018. Annotation and analysis were carried out using EXMARaLDA software tools.

References

de Schryver, Gilles-Maurice. 2010. Getting to the bottom of how language works. In A way with words: Recent advances in lexical theory and analysis: a Festschrift for Patrick Hanks, 3–34. Menha

Publishers.

Hanks, Patrick / El Maarouf, Ismail / Oakes, Michael. 2018 “Flexibility of multiword expressions and corpus pattern analysis.

In Sailer, Manfred / Markantonatou, Stella (eds.) Multiword Expressions: Insights from a Multi-lingual Perspective. Berlín: Language Science Press. 93-119.

McEnery, Tony / Wilson, Andrew 2004. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Shlesinger, Miriam 1998. Corpus-Based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies. *Meta*. 43/4, 486–493.

Using multilingual parallel corpus for Journalistic Translation Research: (re)constructing national images via global news in English, Chinese and Spanish

Biwei Li CT

University of Oviedo

With the globalization of journalistic activities, influential news media are witnessing the increasingly important role played by multilingualism and, consequently, multilingual news translation. In corpus-based translation studies, parallel corpus has proven to be an effective method for contrastive and comparative studies of language (Johansson 1998). However, despite the urgent need for a more comprehensive methodology regarding cross-linguistic comparisons in Journalistic Translation Research or JTR (Valdeón 2020), which is an emerging offshoot within the area of Translation Studies, the employment of parallel corpora (bilingual or multilingual) in this field is still far from abundant, let alone some serious challenges JTR scholars must confront (Caimotto & Gaspari 2018: 209-212).

Under these circumstances, the New York Times Multilingual Parallel Corpus (NYTMPC) is compiled, as part of our PhD project on the construction of national image through news translation. The parallel corpus contains more than 600 news texts in English, Chinese and Spanish published by the American traditional quality media, The New York Times, which has taken on the news production in English as the source language, and their translations into Chinese and Spanish. In our presentation, we will firstly introduce the designing criteria and building procedures of NYTMPC, including the selection, aligning and tagging of the source and target texts. Secondly, by drawing on a combined theoretical framework of imagology (Van Doorslaer 2019) and narrative embedded in translation studies, and the methodologies provided by corpus-assisted discourse studies (Partington et al. 2013) and critical discourse analysis for cross-linguistic comparisons (Baker et al. 2008), preliminary results will be exhibited to illustrate how national images are (re)constructed in multilingual news translation through the (re)shaping of image-related discourses and narratives. Finally, the potentials of parallel corpora for further research on the role played by translation in global news dissemination will also be discussed.

References

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyż anowski, M., McEnery, T., & Wodak, R. (2008). "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press." *Discourse & society*, 19 (3), 273-306.

Cristina Caimotto, M., & Gaspari, F. (2018). "Corpus-based study of news translation: Challenges and possibilities." *Across Languages and Cultures*, 19 (2), 205-220.

Johansson, S. (1998). "On the role of corpora in cross-linguistic research." In *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, edited by S. Johansson and S. Oksefjell, 3-24. Amsterdam-Atlanta, GA: Rodopi.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)* (Vol. 55). Amsterdam/Philadelphia: John Benjamins Publishing.

Valdeón, R. A. (2020). “Journalistic translation research goes global: theoretical and methodological considerations five years on.” *Perspectives*, 28 (3), 325-338.

Van Doorslaer, L. (2019). “Embedding imagology in translation studies.” Слово. ру: балтийский акцент, 10 (3), 56-68.

A Corpus-based Contrastive Study of Verb Tenses and Temporal Adverbs among Spanish, English, and Chinese

Hui-Chuan Lu; An Chung Cheng

CT

National Cheng Kung University; University of Toledo

This study focuses on the co-appearance of verb tenses and temporal adverbs to determine the similarity and distance among Spanish, English, and Chinese by comparing and contrasting the target linguistic combinations in a trilingual parallel corpus of these languages. This investigation, rarely studied previously in the field, is a subset of a larger project on the influences of cross-language transfer on Spanish acquisition by Chinese-speaking learners in Taiwan, whose first foreign language is English (L2) and second foreign language is Spanish (L3). Spanish is considered a morphologically rich language, but Chinese has very poor morphology and English in between. Compared to English, Spanish verbal inflections of tense and aspect are much more complicated. At the same time, Chinese can only express the sentential tense through aspectual markers, temporal adverbs, and contextual cues.

Cross-language contrastive analysis of multilingual expressions with equivalent semantics based on the "Parallel Corpus of Spanish, English and Chinese" was conducted to obtain systematic results for language similarity and difference. Ten top high-frequency Spanish adverbs "*ya*, *hoy*, *ahora*, *luego*, *siempre*, *mientras*, *ayer*, *aún*, *finalmente*, *hasta*" were searched as keywords to extract related data for trilingual analysis.

The results of parallel corpus-based analyses show that the language distinctions in form and meaning connection. For example, Spanish "*ya*" with present tense can correspond to present, present progressive and present perfect tenses in English and adverb "yijing/already" and aspectual marker "le" in Chinese. The results could provide linguistic insights in examining the acquisition of L3 Spanish verbal tenses by Taiwanese learners.

Light verb constructions as a testing ground for the Gravitational Pull Hypothesis: An analysis based on the COVALT corpus

Josep Marco

CT

Universitat Jaume I

The aim of this paper is to test out the Gravitational Pull Hypothesis (GPH) on certain types of light verb constructions (LVC). The study draws on the English-Catalan (EN-CA) and English-Spanish (EN-ES) sub-corpora, and the non-translation components (CA and ES) in COVALT – originally a multilingual corpus made up of the translations into Catalan of narrative works originally written in English, French, and German published in the autonomous region of Valencia from 1990 to 2000, together with their corresponding source texts. Spanish translations of the same source texts as well as two comparable components of Catalan and Spanish non-translations were later added. The translated components of EN-CA and EN-ES comprise 1,343,631 and 1,122,299 words, respectively. As to the non-translated components, CA is made up of 1,551,521 and ES of 4,170,178 words.

The GPH was put forward by Halverson (2003) as an attempt to bring together various alleged properties of translated text, such as over- and under-representation of target language typical features. It posits three potential causes of translational effects: patterns of salience or prototypicality, which are target language internal (factor 1); conceptual structures/representation of the source language item, which are related to the structure of the source language (factor 2); and patterns of connectivity, which reflect relationships between the source and the target languages (factor 3). One effect is predicted for each potential cause, or factor. The effect of factor 1 will be over-representation; the effect of factor 2 will be over-representation too; and the effect of factor 3 may be over- or under-representation.

A LVC is made up of a predicative noun (i.e. a noun with an internal argument structure which can therefore assign roles) and a light or support verb (i.e. a partly demanticised verb which contributes little or no meaning to the construction and serves to root the predicate in time and assign the roles inherent in the predicate). LVCs have given rise to a considerable body of literature that cannot be summarised here. Only two aspects will be briefly touched upon because they will be operationalised in the research reported. Firstly, the limits of the notion are not clear. I will adhere to the two restrictions posited by Colominas (2001), according to which a LVC is a complex predicate whose meaning can be compositionally derived. The compositionality restriction sets LVCs apart from other, more prototypical phraseological units. In that respect, LVCs are a kind of collocation. And secondly, predicative nouns, insofar as they designate a situation or state of affairs and are able to assign semantic roles, can be semantically classified along the same lines as verbs. Many different classifications have been put forward, all of them more or less indebted to Vendler's (1967) original one. I will follow Alonso Ramos (2004), who identifies seven categories: state, quality, action, activity, act, process, and event. The two criteria underlying this classification are volition, or agentivity, and duration.

LVCs have been chosen as a testing ground for the GPH because the degree of isomorphism both in semantic and syntactic terms between a target language LVC and its corresponding source text trigger can crucially impact its frequency in translated as opposed to non-translated texts via their patterns of connectivity. The other two factors at work in the GPH (target language and source language salience) being difficult to establish in the present case, only the factor of connectivity will be taken

into account.

The method employed in this study can be summarised as follows:

a. searching for main predicative nouns co-occurring with *fer* in CA and *dar* in ES

b. identifying meaning patterns in the ensuing LVCs

c. formulating hypotheses

d. searching for main predicative nouns co-occurring with *fer* in EN-CA and *dar* in EN-ES

e. determining similarities and differences in order to test out the hypotheses

f. identifying ST triggers for Catalan and Spanish LVCs in EN-CA and EN-ES, respectively, in order to account for similarities and differences

Both *fer* in Catalan and *dar* in Spanish are prototypical light verbs in that they have a very low level of semantic specificity and enter into a large number of LVCs. One of the patterns identified in the data retrieved from Catalan and Spanish non-translations is that they often co-occur with such nouns as *por/miedo* ('fear') or *llàstima/lástima* ('pity') to convey emotional states. The degree of isomorphism between these constructions and their potential English triggers (as attested by bilingual dictionaries) is lower than that between LVCs designating acts or actions (e.g. *fer una ullada* 'have/take a look' or *dar un beso* 'give a kiss') and their potential English triggers. That may result in a lower degree of connectivity in the former group of LVCs, leading to under-representation in translated texts.

On the basis of the above, the following hypotheses are formulated:

1. LVCs based on *fer* denoting emotional states will be under-represented in EN-CA as compared to Catalan non-translations.

2. LVCs based on *dar* denoting emotional states will be under-represented in EN-ES as compared to Spanish non-translations.

3. LVCs based on *fer* denoting actions will show no significant frequency differences in EN-CA and Catalan non-translations.

4. LVC based on *dar* denoting actions will show no significant frequency differences in EN-ES and Spanish non-translations.

Preliminary results confirm hypotheses 1 and 2 in most cases, for such LVCs as *fer gràcia* ('fun') / *por* ('fear') / *ràbia* ('anger') in Catalan and *dar miedo* ('fear') / *vergüenza* ('shame') / *asco* ('disgust') in Spanish. The case with hypotheses 3 and 4 is much less clear-cut. Over and above hypothesis validation, the analysis of degrees of connectivity between target language LVCs and their source text triggers is expected to have explanatory value.

References

Halverson, Sandra. 2003. “The cognitive basis of translation universals”. *Target* 15(2): 197–241.

Vendler, Zeno. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.

Colominas, Carme. 2001. *La representació semàntica de les construccions de suport des d’una perspectiva multilingual*. Bellaterra: Universitat Autònoma de Barcelona (PhD dissertation).

Alonso Ramos, Margarita. 2004. *Las construcciones con verbo de apoyo*. Madrid: Visor.

The Chinese-Spanish Corpus of Journey to the West

Tian Mi; Rodrigo Muñoz

CT

Nankai University

The built-up of a parallel corpus using a Chinese novel that was written eight hundred years ago and its two complete translations carried out into Spanish. The scope of this corpus is to carry out multidisciplinary researchers, as translation, historic, literary, paremiology and contrastive grammar studies.

As regards its compilation, the first step was to digitalize the texts, not an easy task because each of these texts consists of nearly a thousand pages each. Regretfully, we were unable to receive from the publishing houses a copy of them in digital form, so we were obliged to do the task manually. The second step was to align them. We used Microsoft Excel xlsx format files to align the book at the sentence level, using three parallel cells to place the source text, target text 1 and target text 2. Every sentence was numbered and the work stored in one hundred files; one for each chapter.

Researches on the CCEVAO corpus started one and a half years ago. The first study was centered on the different meaning numbers have in these two cultures. We selected number nine and our job was finally published in the journal Estudios de Traducción last December, titled "¿Numerales que atraviesan la barrera lingüística? Un estudio práctico de la traducción del 9 del chino al español por medio del corpus CCEVAO".

The second study was focused on analyzing the translation of proverbs. Our intention was two-fold: On one hand, we aimed at how sayings were translated and which strategies were used to transfer their meanings to the target language. On the other hand, we were interested in determining how sayings that were written eight hundred years ago were translated into today's Spanish and which sayings -old, new or invented- were used in the process. This job was finally published last December in the journal Paremia last December, being titled "Translación de los *suyan* chinos al español: un nuevo reto".

Three further works are being in the peer-reviewing process:

"Análisis comparativo de las traducciones de los títulos de los capítulos de *Viaje al Oeste* de Chino a español" was focused to study the translation of the headings of the one hundred chapters of the novel, paying special attention to the differences found in the three texts at the syntactic, semantic and lexical levels.

"El humor de los juegos de palabras en Viaje al Oeste, ¿un viajero a bordo?" was conceived as a research on humour, which is deeply influenced by culture and religion. Likewise, we also focused on how the irony of the wordplays and the characters with double meanings was translated into Spanish. In other words: how translators negotiated linguistic barriers to translate their real meaning.

"Recreación del mundo fantástico interpersonal: traducción de los antropónimos en *Viaje al Oeste*". In this case, anthroponomy denotes people's individual qualities affected to a specific culture and morale, and also embodies a varied range of extralinguistic information and an aesthetic symbolism absent in Spanish.

References

Attardo, S. (2001). Humorous Texts: A Semantic and Pragmatic Analysis. New York, É.- U.: Mouton de Gruyter.

Cao, S., da Cunha, I., & Iruskieta, M. (2016). A Corpus-based Approach for Spanish- Chinese Language Learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)* (pp. 97-106).

Cao, S., & Gete, H. (2018). Using Discourse Information for Education with a Spanish- Chinese Parallel Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Corpas Pastor, G. (1996). Manual de fraseología española, Madrid, Ed. *Gredos (Biblioteca Románica Hispánica III. Manuales, 76)*.

Li, Q., «Moshi shuzi jiu de wenhua chanshi» («Interpretación cultural del numeral de modelo, el 9»), *Revista de la Universidad de las Nacionalidades de Guangxi* (2003), 68-69.

Li, X. (2012). *Zhonguo gudianxiaoshuo huimu yanjiu (Estudios sobre los títulos de capítulo de las novelas clásicas chinas)*, Pekín: Peking University Press.

Lou, G. (1985). Cong Xingming kan shehui he wenhua (Observación sobre la sociedad y la cultura mediante los nombres). *Foreign Language Teaching and Research*, 63, p. 14-19.

Nida, E.A. (2017): *Language and Culture*. Contexts in Translating, Shanghai: Shanghai Foreign Language Education Press.

Ku, M. (2019). Viaje al Oeste vs. viaje a la diversión: estrategias de traducción de los elementos culturales de Peregrinación al Oeste. *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*, (43), 50-69.

Raskin, V. (1985). Semantic Mechanisms of Humor. Dordrecht Boston Lancaster: D.

Resnik, P., Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3), 349-380.

Scheu, Ú. D. (1996). Creencias y mitos en el uso del número en tres culturas europeas. *Revista murciana de antropología*, (3), 61-70.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *arXiv preprint cs/0609058*.

St John, E. (2001). A Case for Using a Parallel Corpus and Concordancer for Beginners of a Foreign Language. *Language learning technology*, 5(3), 185-203.

Tanaka, Y. (2001). Compilation of a Multilingual Parallel Corpus. *Proceedings of PACLING 2001*, 265-268.

Volk, M., Graën, J., & Callegaro, E. (2014). Innovations in Parallel Corpus Search Tools.

Wang, J. (1990). *Wenhua de jingxiang——renming (Reflejo de la cultura: los antropónimos)*. Changchun: Educación de Jilin.

—Es el fin —balbució—, *„Das ist das Ende“, stammelte er*. Reporting direct speech in Spanish and German

Teresa Molés-Cases CT

Universitat Politècnica de València

This contribution examines manner-of-speaking in a Spanish>German parallel corpus of narrative texts extracted from the online corpus PaGeS (Parallel Corpus German Spanish, Universidade de Santiago de Compostela). The inspiration for this research has been the observation that manner-of-speaking has not yet been paid due attention in the Thinking-for-translating framework (Slobin 1996, 2000), in comparison with the phenomenon of manner-of-motion (exceptions mainly include Caballero 2015, Rojo and Valenzuela 2001 and Shi 2008). According to Slobin's Thinking-for-translating hypothesis, translators tend to distance themselves from the source text in order to conform to the rhetorical style of the target language (usually their mother tongue). One of the translation problems most commonly analyzed in translation scenarios including languages belonging to different typologies is manner-of-motion (e.g. Cifuentes-Férez 2006, 2013; Filipović 1999, 2008; Ibarretxe-Antuñano 2003; Lewandowski and Mateu 2016). The data so far indicate that, in translation from a verb-framed language (e.g. Spanish) into a satellite-framed language (e.g. German), addition of manner-of-motion is usually observed (e.g. [...] y luego se dirigió al árbol. → [...] *und dann schlenderte er auf den Baum zu*.|"[...] and then he strolled to the tree"), whereas in the inverse typological combination, there is a tendency for this component to be omitted (e.g. *Dann flog sie ganz dicht an sein Ohr heran* [...]|"Then she flew very close to his ear [...]" → Luego, acercándose mucho a su oído [...]).

The aim of this contribution is twofold: first, to examine translators' behavior regarding manner-of-speaking in a verb-framed language > satellite-framed language translation scenario (Spanish>German), specifically focusing on the translation of reporting verbs introducing direct speech in a corpus of narrative texts; second, to compare the resulting data with findings from previous comparable studies dealing with the communication and motion frames. Here we resort to a classification of the examined verbs into two categories, the Spanish general verb *decir* and a series of manner-of-speaking verbs (i.e. a classification of 55 manner verbs of human speech included in Caballero 2015, e.g. *balbucear*, *murmurar*). This study makes use of a section of the online corpus PaGeS, and more specifically, it analyzes 10 contemporary novels originally written in Spanish and the corresponding translations into German. The results suggest that translators' behavior differs when dealing with the motion and communication frames: while manner-of-motion is often added in translations into German (from Spanish), this tendency is observed in few cases in the case of manner-of-speaking (e.g. —¿No me ha oído? —dijo el operador—. → *„Haben Sie mich nicht verstanden?“, knurrte der Kameramann*.|"Didn't you understand me?" the cameraman growled.). However, inclusion of other frame elements has been identified in the analysis of the general verb *decir*: for instance, intention (—¿Tarde? — dije. → *„Spät?“, fragte ich*.|"Late?" I asked.). The resulting data from the analysis of manner-ofspeaking verbs indicate that manner is mostly maintained (—Es el fin —balbució—. → *„Das ist das Ende“, stammelte er*.|"This is the end," he stammered.), with only a few cases of omission (—[...] seguro —se echó a reír [...] → *„[...] sicher“, antwortete er* [...]|"[...]sure," he answered).

Keywords: manner-of-speaking, reporting verbs, narrative texts, direct speech, Thinking-for-translating.

References

Caballero, Rosario (2015). Reconstructing speech events: Comparing English and Spanish. *Linguistics*, 53(6), 1391–1461.

Cifuentes-Férez, Paula (2006). *La expresión de los dominios de movimiento y visión en inglés y en español desde la perspectiva de la lingüística cognitiva*. Universidad de Murcia: Master Thesis.

Cifuentes-Férez, Paula (2013). El tratamiento de los verbos de manera de movimiento y de los caminos en la traducción inglés-español de textos narrativos. *Miscelánea: A Journal of English and American Studies*, 47, 53–80.

Filipović, Luna (1999). *Language-specific expression of motion and its use in narrative texts*. University of Cambridge: PhD thesis.

Filipović, Luna (2008). Typology in action: applying typological insights in the study of translation. *International Journal of Applied Linguistics*, 18(1), 23–40.

Ibarretxe-Antuñano, Iraide (2003). What Translation tells us about Motion: A Contrastive Study of Typologically different Languages. *IJES, International Journal of English Studies*, 3(2), 151–175.

Lewandowski, Wojciech and Mateu, Jaume (2016). Thinking for translating and intratypological variation in satellite-framed languages. *Review of Cognitive Linguistics*, 14(1), 185–208.

PaGeS (Parallel Corpus German Spanish), Universidade de Santiago de Compostela, https://www.corpuspages.eu/, (01.02.2021).

Rojo, Ana and Valenzuela, Javier (2001). How to Say Things with Words: Ways of Saying in English and Spanish. *Meta: Journal Des Traducteurs*, 46(3), 467–477.

Shi, D. (2008). Communication verbs in Chinese and English: A contrastive analysis. *Languages in Contrast*, 8(2), 181–207.

Slobin, D. (1996). Two ways to travel: Verbs of motion in English and Spanish. In M. Shibatani & S. A. Thompson (Eds.), *Grammatical Constructions: their form and meaning* (195–220). Oxford: Clarendon Press.

Slobin, Dan (2000). Verbalized events: A dynamic approach to linguistic relativity and determinism. In S. Niemeier & R. Dirven (Eds.), *Evidence for linguistic relativity* (107– 138). Berlin: Mouton de Gruyter.

Las herramientas semiautomáticas de redacción y traducción español-inglés basadas en corpus: el ejemplo de GEFEM

María Teresa Ortego Antón CT

CITTAC, Universidad de Valladolid

La internacionalización del sector agroalimentario en España ha propiciado un aumento exponencial de los servicios de redacción y de traducción del español al inglés. En este contexto socioeconómico, describimos la metodología empleada para construir GEFEM, una herramienta basada en corpus que asiste durante la redacción y traducción del español al inglés de un determinado género textual: las fichas descriptivas de embutidos. En primer lugar, se ha compilado, anotado y explotado un corpus virtual comparable en español y en inglés (C-GEFEM) compuesto por fichas descriptivas de embutidos siguiendo el protocolo propuesto por Seghiri (2017) y Ortego Antón (2019). El resultado es un corpus virtual comparable bilingüe compuesto por 100 textos redactados originalmente en cada lengua, con un tamaño de 14196 tipos en español y 24604 en inglés. La diferencia de tamaño entre lenguas se debe a que las fichas descriptivas de embutidos en lengua inglesa especifican la información del producto, las características de empaquetado o el uso, en tanto que en español son mucho más simples.

Una vez compilado el corpus, hemos extraído y analizado los movimientos y pasos (Biber *et al.*, 2007: 23-24) que constituyen la estructura retórica en español y en inglés (Ortego Antón, 2020) etiquetando los subcorpus en español y en inglés de C-GEFEM con la ayuda del Etiquetador de Movimientos Retóricos©[3] desarrollado por el grupo interuniversitario ACTRES[4]. Con el Visor de Corpus Comparables©[5] establecemos la estructura retórica prototípica en cada lengua. Además, se ofrecen los patrones léxico gramaticales más recurrentes para un determinado movimiento o paso a través de las líneas modelo, definidas estas últimas como las oraciones típicas donde el contenido y el formato es estándar (López Arroyo y Roberts, 2015: 157). Las líneas modelo se presentan con huecos que los usuarios pueden completar y están enriquecidas con un glosario bilingüe español-inglés con siete categorías: aditivos, alérgenos, elementos nutricionales, empaquetado, ingredientes, materiales, origen y país. Por último, con estos datos se desarrolló GEFEM, el generador semiautomático de fichas descriptivas de embutidos, que incluye la estructura retórica con los movimientos y pasos más frecuentes, las principales líneas modelo y sus patrones léxico gramaticales y un glosario terminológico y fraseológico bilingüe (Ortego Antón, 2021).

Por tanto, GEFEM puede considerarse una herramienta de traducción fiable, basada en corpus y fácil de utilizar que asistirá a traductores y redactores a trasvasar los textos de un determinado género textual, las fichas descriptivas de producto del español al inglés, de manera que esta herramienta semiautomática ofrezca respuesta a las demandas de trasvase interlingüístico del español al inglés del sector agroalimentario durante la expansión internacional de sus ventas.

Palabras clave: traducción, corpus comparable, herramienta semiautomática, inglés, español.

Referencias

Biber, Douglas; Connor, Ulla y Thomas A. Upton. 2007. *Discourse on the Move. Using Corpus Analysis*

[3] http://contraste2.unileon.es/web/en/tagger.html (Consulta: 30/01/2021).
[4] https://actres.unileon.es/wordpress/?lang=en (Consulta: 30/01/2021).
[5] http://contraste2.unileon.es/web/es/browser.html (Consulta: 30/01/2021).

to Describe Discourse Structure. Amsterdam: John Benjamins. DOI:10.1075/scl.28

López Arroyo, Belén y Roda P. Roberts. 2015. "The use of comparable corpus: How to develop writing applications". En María Teresa Sánchez Nieto (Ed.), *Corpus-based Translation and Interpreting Studies: From description to application / Estudios traductológicos basados en corpus: de la descripción a la aplicación*. Berlin: Frank und Timme, 147-165.

Ortego Antón, María Teresa. 2019. *La terminología del sector agroalimentario (español-inglés) en los estudios contrastivos y de traducción especializada basados en corpus: los embutidos*, Berlin: Peter Lang.

Ortego Antón, María Teresa. 2020. «Las fichas descriptivas de embutidos en español y en inglés: Un análisis contrastivo de la estructura retórica basado en corpus», *Revista Signos. Estudios de Lingüística*, vol. 53, n. 102, pp. 170-194 [DOI: 10.4067/S0718-09342020000100170].

Ortego Antón, María Teresa. 2021. A Spanish-English frame-based e-dictionary about dried meats. *Terminology*. DOI: 10.1075/term.20013.ort

Seghiri, Míriam. 2017. "Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores". *Babel* 63 (1): 43-64. DOI: 10.1075/babel.63.1.04seg

Looking at Flemish tapestries from the wrong side: Translation (Spanish-English) of cultural references in rural tourism hospitality industry web pages

Leonor Pérez-Ruiz CT

University of Valladolid

Spanish Rural tourism accommodation Web pages, translated into English, are a great means to attract international visitors to the area. Being closely related to culture, the language of tourism contains a great amount of culturemes. These culturems have been recently studied -based on seminal works by Vermeer (1983), House (1997), Katan (1999), Franco Aixelá (1996), Molina Martínez (2001), among others- by various authors (Comitre Narváez, 2004; Terestyényi, 2011; González Pastor, 2012; Soto Almela, 2013 2014) and are considered "a potential source of untranslability" (Comitre Narváez & Valverde Zambrana, 2014: 71) due to the difficulty that describing the characteristic elements of a culture poses.

Being part of a wider study, the aim of this paper has been to verify the quality, precision and clarity of the translation of cultural references in Castile and Leon Rural tourism accommodation Web pages, as well as to identify the main translation techniques applied. For the purpose of this study, following Franco Aixelà, we have considered those culturems "whose function and connotations in a source text involve a translation problem in their transference to a target text" (1996:58). Also we have analyzed those culturems pertaining the rural tourism accommodation itself, i.e. buildings, rooms, facilities, furniture and appliances, and food.

Data have been drawn from the exploitation of an ad hoc parallel corpus (Spanish-English) made up of texts collected from Rural Tourism hostels in the autonomous region of Castile and Leon. With the aim of analyzing if the usage, translation and/or omission of cultural references varies noticeably among regions which have a high or low rates of international tourism, we compared these results with those from two other autonomous regions corpora (Andalusia and Valencia).

The corpora compiled contain 166 texts totaling 190,907 words (102,191 in Spanish and 88,716 in English). The identification of the different culturally loaded terms in Spanish was done by using AntConc 3.5.7 corpus analysis tool software (Anthony, 2018), generating Word Lists that were later analyzed manually for the extraction of the culturems. The terms were later arranged in different lists according to the corresponding semantic fields. The corpora had been previously aligned manually and thus the corresponding translations were identified in the English corpora and categorized following Gonzalez Pastor (2018) classification. Finally, they were evaluated.

We found that 47% of the culturemes identified were used indistinctly in all the regions (e.g. bodega, chorizo or cabaña), and their translation tends to follow a similar pattern. The rest were specific of one of these territories or were adapted somewhat, as in the case of the term cocido translated as Castilian stew. As for the translation techniques more frequently used, direct translation (village), borrowing (pueblo) and generalization (mushroom for boletus) were the commonest. Also, the translation of some terms was not consistent, and different options were used depending on the case, e.g. patio was translated as courtyard, yard or garden, and the borrowing patio was also used in the target text. Some of these translations were not adequately used given the context. Also, frequent cases of omission have been identified (plaza was omitted in 72 instances in the Andalusian sub-corpus and sierra not translated in 31 instances in the Castilian one).

We conclude that, as expected, still much can be done in order to avoid incorrect, not accurate or omitted information in these Webs. All these faults often provoke misleading information or misunderstandings. Finally, given the abundance of cultural-bound terms in the texts analyzed, we believe this study will contribute to better promote Spanish rural tourism and introduce our culture to foreign tourists.

References

Anthony, L. (2018). AntConc (Versión 3.5.7.) [Computer Software]. Tokyo, Japan: Waseda University.

Comitré Narváez, I. (2004). "La traducción de culturemas en publicaciones del sector turístico. Un estudio empírico", en Gallegos Rosillo, J.A. y Benz Busch, H. (eds.), Traducción y cultura: el papel de la cultura en la comprensión del texto original. Málaga: Encasa, 115-138.

Comitre Narváez, I.., & Valverde Zambrana, J. M. (2014). How to translate culture-specific items: a case study of tourist promotion campaign by Turespaña. The Journal of Specialised Translation, 21, 71-112.

De la Cruz Trainor, M. M. (2004). "Traducción al inglés de términos culturales en textos turísticos", en Gallegos Rosillo, J.A. y Benz Busch, H. (eds.), Traducción y cultura: el papel de la cultura en la comprensión del texto original. Málaga: Encasa, 83-114.

Dörnyei, Zoltan. 2007. Research Methods in Applied Linguistics. Oxford: Oxford University Press. Franco Aixelà, J.(1996). "Culture-specific ítems in Translation." In Álvarez, R. and Vidal, A. (eds) (1996). Translation, Power, Subversion. Clevedon: Multilingual Matters, 52-78.

González Pastor, D. M. (2012): Análisis descriptivo de la traducción de culturemas en el texto turístico. Tesis doctoral. Universitat Politècnica de Valencia.

House, J. (1997). Translation quality assessment: A model revisited. Gunter Narr Verlag.

Katan, D. (1999). Translating Culture, an Introduction for Translators, Interpreters and Mediators. Manchester: St. Jerome Publishing.

Molina Martínez, L. (2001). Análisis descriptivo de la traducción de los culturemas árabe-español. Tesis doctoral, Universidat Autònoma de Barcelona.

Pamies, A. (2017). The concept of cultureme from a lexicographical point of view. Open Linguistics, 3(1), 100-114.

Payo Peña, L. (2002). "La traducción de las referencias culturales en un texto turístico". Puentes: hacia nuevas investigaciones en la mediación intercultural, 1, 1, 33-45.

Soto Almela, J. (2013). La traducción de culturemas en el ámbito del patrimonio cultural: análisis de folletos turísticos de la Región de Murcia. Revista de estudios filológicos, 24.

Soto Almela, J. (2014). Los Términos Culturales en el Ámbito Turístico EspañolInglés: Traducción, Manipulación y Recepción Real en Usuarios Anglófonos. Tesis doctoral no publicada, Universidad de

Murcia, España.

Terestyényi, E. (2011). Translating culture-specific items in tourism brochures. SKASE journal of translation and interpretation, 5(2), 13-22.

Vermeer, H. (1983). “Translation theory and linguistics”. In Roinila, P., Orfanos, R. Tirkkonen- Condit, S. (Eds.), Näkökohtia käänämisen tutkimuksesta (1-10). University of Joensuu.

Light Verb constructions in translation. What corpora tell us

Rosa Rabadán Álvarez

University of León

Light verb constructions (LVCs) are complex predicates involving a verb of general meaning (the so-called light verb) like make/*hacer* or take/*tomar* plus an abstract, generally eventive Noun Phrase, which may also include a modifier, often an article (Alonso Ramos 2004, Butt 2010). Grammatically, LVCs are just like any other V + NP pattern; the light verb is considered weak, serving only as a verbal marker with most of the meaning being supplied by the NP (Leech, 2006; Spencer, 2013). LVCs can be rephrased and seem to be equivalent to their correlated verb, i.e., make a decision> decide, take a rest> rest (Nenonen et al. 2017). Lexically, LVCs are often considered multiword expressions (MWE) and treated as a fixed dictionary collocation. However, only some combinatorial dictionaries list them consistently as LVCs (e. g., Sp REDES). Databases such as UCREL's USAS do so haphazardly, e. g. 'make a mistake' make_A5.3-[i1.2.1 a_Z5 mistake_A5.3-[i1.2.2 is marked as an MWE, whereas 'make a decision [make_A1.1.1 a_Z5 decision_X6+] is treated as a list of independent elements. Recent research indicates that LVCs constitute a semi-productive category, not a closed list of, and efforts are devoted to establishing semantic compatibility features that underlie LVCs (De Miguel 2011). Empirical data also show that LVCs may convey additional, more specialized grammatical meanings that are not present in the single correlated verb, such as aspectual, volitional, or causative notions (Sanromán Vilas 2017).

This paper reports on WiP research and explores how LVCs are conveyed in translation into Spanish. The aim is to discover how these more specialized meanings are rendered into Spanish and evaluate the different translation solutions' consequences.

To do this, we will use three corpora: P-ACTRES 2.0, an English-Spanish parallel corpus including originals and translations in both directions; CETRI, a corpus of Spanish translated from English, and CORPES XXI (v. 0.93), the monitor corpus of contemporary Spanish.

Parallel corpus data show that LVCs follow regular translation routines, formal solutions, lexical reinterpretation and, frequently, translation by the single correlated verb. This suggests that users are aware of the constructions' unitary meaning but entirely unaware of their meaning specialization. CETRI and CORPES XXI provide quantitative findings to establish possible differences between translated and non-translated data For the pair English-Spanish, preliminary results indicate that these constructions' specificity is ignored mainly in translation, which results in changes in formality scale and, less frequently, in meaning processing and interpretation. This state of affairs suggests it would be beneficial for contrastive and translation research to encode LVCs information in annotation routines, either as part of already existing layers or in the form of an additional grammatical meaning layer (Nagy et al., 2011, Vincze et al. 2011 2013). This would also contribute to making better translation decisions in HT and post-editing or assigning high-quality MT or CAT translation options.

Keywords: Light verb constructions, parallel corpus, grammatical meaning, translation solutions.

References

Alonso Ramos, M. 2004. *Las construcciones con verbos de apoyo*, Madrid: Visor.

Bosque, Ignacio (dir.), 2004. REDES. *Diccionario combinatorio del español contemporáneo*. Madrid: SM.

Butt, M. 2010. The light verb jungle: Still hacking away. In M. Amberber, B. Baker, M. Harvey (Eds.), *Complex Predicates: Cross-linguistic Perspectives on Event Structure* (pp. 48-78). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511712234.004

De Miguel, E. 2011. En qué consiste ser verbo de apoyo. In M. V. Escandell Vidal, M. Leonetti, C. Sánchez López (eds.), *60 problemas de gramática dedicados a Ignacio Bosque*. Madrid: Akal, 139-147.

Leech, Geoffrey. 2006. *A Glossary of English Grammar*. Edinburgh University Press.

Nenonen, Marja Juha Mulli, Alexandre Nikolaev and Esa Penttilä. 2017. How light can a light verb be? Predication patterns in V + NP constructions. In Milla Luodonpää-Manni, Esa Penttilä, Johanna Viimaranta (eds), *Empirical Approaches to Cognitive Linguistics: Analyzing Real-Life Data*. Newcastle: Cambridge Scholars Publishing.

Sanromán Vilas, Begoña . 2017. From the heavy to the light verb: An analysis of TOMAR 'to take.' *Lingvistica Investigationes* 40(2).

Spencer, Andrew. 2013. *Lexical Relatedness: A Paradigm-Based Model*. Oxford University Press.

Exploring near-synonyms through translation corpora: A case study on *begin* and *start* in the English-Spanish Parallel Corpus PACTRES

Noelia Ramón CT

University of León

This paper focuses on the translation solutions found in Spanish for two aspectual verbs in English which are often considered to be near-synonyms: *begin* and *start*. These two verbs indicate an ingressive aspect and occur in similar syntactic constructions, including to-infinitive complementation, -ing clause complementation or nominal complementation. However, previous studies have shown that *begin* and *start* appear to have different semantic preferences to some extent and are consequently not completely synonymous (Freed 1979; Dixon 1991, 2005; Duffley 1999; Mair 2002; Egan 2008). For example, it has been found that *start* is significantly more common in spoken registers of English than *begin* (Egan 2008: 257) and "only from a sentence with *begin* does it necessarily follow that the nucleus (or characteristic activity) of the event has been initiated; a sentence with *start* followed by a to V complement can have as a consequence that only the onset of the event named in the complement has been initiated." (Freed 1979: 71).

In this study we will analyse the various translational options provided in Spanish for the verbs *begin* and *start* in order to determine if the semantic differences are reflected in the translators' choices. The empirical data for this study have been extracted from the PACTRES parallel corpus of English and Spanish compiled at the University of León, Spain (Sanjurjo-González Izquierdo 2019). PACTRES is a bidirectional parallel corpus, but for the present study we have focused exclusively on English originals and their translations into Spanish. All the texts included in the corpus have been published in the year 2000 or later and belong to different registers such as fiction, essays, periodical publications or miscellanea. PACTRES contains a total of 1,551 concordance lines of the verb *begin* and 981 of the verb *start*, including all their morphological forms, with their corresponding translations into Spanish. A sufficiently representative sample was selected to carry out the analysis of the translational solutions.

The results of the study show a clear dominance of the lexical verb *empezar* in Spanish to convey the meaning of both English verbs. Additional translational options include other ingressive verbs such as *comenzar*, *iniciar*, *ponerse a*, etc., as well as strategies like omission or modulation. The differences found in the Spanish translations can be used as an indicator of the actual semantic nuances distinguishing the uses of *begin* and *start* in the source language English. Translation corpora "contain the intuitive linguistic responses of competent language users to a series of linguistic prompts" (Egan 2012: 13), so the translational solutions provided highlight differences in meanings between near-synonyms in the source language.

The aim of this paper is to establish the inventory of translational options available in Spanish to cover the semantic field of ingressive aspect expressed lexically by the verbs *begin* and *start* in English. In addition, the results confirm that data from parallel corpora contribute valuable information with regard to semantic differences between near-synonymous lexical items.

References

Dixon, R.M.W. 1991. *A New Approach to English Grammar on Semantic Principles*. Oxford: Clarendon

Press.

Dixon, R.M.W. 2005 *A Semantic Approach to English Grammar*. (2nd edition) Oxford: Oxford University Press.

Duffley, P. 1999. The use of the infinitive and the -ing after verbs denoting the beginning, middle and end of an event. *Folia Linguistica* 33: 295-331.

Egan, T. 2008.*Non-finite Complementation : A Usage-based Study of Infinitive and -ing Clauses in English*. Amsterdam: Rodopi.

Egan, T. 2012. Using translation corpora to explore synonymy and polysemy. *Varieng*. *Studies in Variation, Contacts and Change in English*. Volume 12. Aspects of corpus linguistics: compilation, annotation, analysis. Electronic edition. 1-14.

Freed, A. 1979. *The Semantics of English Aspectual Complementation*. Dordrecht: Reidel.

Mair, C-. 2003. Gerundial complements after begin and start: Grammatical and sociolinguistic factors, and how they work against each other. In Mondorf, B. and G. Rohdenburg (eds.) *Determinants of Grammatical Variation in English*. Berlin: Mouton. 329-343.

Sanjurjo-González, H. Izquierdo, M. 2019. PACTRES 2.0. A parallel corpus for cross-linguistic research. In: I. Doval M. Sánchez Nieto (eds.) *Parallel Corpora for Contrastive and Translation Studies*. Amsterdam/Philadelphia: John Benjamins. 215-232.

Elaboración de un corpus paralelo sobre movimiento causado en las lenguas romances

Daniel Rojas-Plata CT

Tecnológico Nacional de México/Cenidet

En esta presentación, proponemos la elaboración de un corpus paralelo trilingüe (españolfrancés-italiano) que permita la extracción de construcciones de movimiento causado. Nuestro objetivo es presentar la metodología para la formación del corpus y las herramientas que utilizamos para su interrogación.

Las construcciones de movimiento causado han sido descritas por diversos autores de manera teórica (Talmy 1985, 2000; Goldberg 1995) así como empírica (Kopecka Narasimhan, 2012; Ibarretxe-Antuñano, 2017). Sin embargo, salvo el trabajo de Levshina (2015) quien analiza las construcciones causativas en un corpus de subtítulos de películas, no ha sido del todo abordado un análisis contrastivo del movimiento causado desde una perspectiva basada en corpus. La utilidad de un estudio de este tipo concierne tanto a la traducción automática como a la semántica lexical, ya que puede ofrecer datos relevantes sobre el modo en que las diferentes lenguas abordadas expresan un mismo evento de movimiento causado.

A continuación, describimos la metodología utilizada en este trabajo. Debido a la naturaleza de las construcciones analizadas, las fuentes de información utilizadas para la elaboración del corpus son subtítulos de películas y de series televisivas, los cuales obtuvimos del corpus OPUS (Tiedemann Nygaard, 2004). Hemos limitado el número de palabras contenidas en los subcorpus de cada lengua a 10 millones para que su procesamiento y análisis pueda resultar más ágil. El método que utilizamos para extraer y analizar las construcciones de movimiento causado se basa en tres grandes etapas. En la primera, realizamos la búsqueda de un conjunto de verbos de movimiento causado previamente seleccionados en las tres lenguas. La intención de tomar en cuenta sólo algunos verbos reposa en la necesidad de mantener un control sobre los resultados en esta primera fase de la investigación. Como segundo paso, operamos un análisis automático de dependencias de las ocurrencias obtenidas con ayuda de la herramienta FreeLing (Padró, 2011). Esto nos permite establecer cuáles son los argumentos que dependen del verbo en cuestión. Para la última etapa, hemos desarrollado un algoritmo que nos permite seleccionar aquellos argumentos espaciales que se relacionan con el verbo. Así podemos separar aquellas construcciones que expresan un evento espacial de aquellas que no lo hacen.

Este método nos ha permitido obtener datos consistentes sobre el empleo de los verbos de movimiento causado, así como sus respectivas traducciones en las tres lenguas estudiadas. La siguiente fase prevista para el desarrollo de este corpus es la unificación de las diferentes etapas en un solo proceso automatizado.

Referencias

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago/London: University of Chicago Press.

Ibarretxe-Antuñano, I. (ed.). (2017). *Motion and space across languages: Theory and applications* (Vol. 59). Amsterdam/Philadelphia: John Benjamins Publishing.

Kopecka, A., Narasimhan, B. (eds.). (2012). *Events of putting and taking: A crosslinguistic perspective* (Vol. 100). Amsterdam/Philadelphia: John Benjamins Publishing.

Levshina, N. (2015). European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica*, 49(2), 487-520.

Padró, L. (2011). Analizadores Multilingües en FreeLing. *Linguamatica*, 3(2), 13-20.

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

Talmy, L. (2000). *Towards a Cognitive Semantics II: Typology and Process in Concept Structuring*. Cambridge: MIT Press.

Tiedemann, J., Nygaard, L. (2004) The OPUS corpus - parallel free. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal, May 26-28.

Expresión de la perspectiva del recipiente en la traducción alemán-español. Un estudio de las pasivas de *bekommen*, de *erhalten* y de *kriegen* en los datos del corpus PaGeS

María Teresa Sánchez Nieto CT

Universidad de Valladolid

En Sánchez Nieto (2017) estudiamos las correspondencias de la pasiva de bekommen o bekommen-Passiv (Eggelte 2020, 150–51; Drosdowski 1995, 178; Engel 1996, 454) en español surgidos en situaciones de mediación lingüística. Para ello, extrajimos del corpus Europarl ejemplos de las construcciones de pasiva de bekommen y sus correspondientes alineaciones en español. Sobre la base de estos ejemplos, estudiamos a) hasta qué punto se mantenía la perspectiva del recipiente (beneficiario o maleficiario) con la descentralización del agente propias de la pasiva de bekommen en la organización de la información dentro de la oración y b) el coste léxico que emanaba como resultado de comparar el predicado formulado con la pasiva de bekommen en la oración alemana y el predicado de la oración española correspondiente. Del total de 281 ejemplos de pasiva de bekommen analizados, en el 54,8 % se había recurrido a la activa española, en el 27,8 % a la pasiva refleja, en el 6 % a una nominalización que recogía el significado del verbo principal, en el 3,6 % a la pasiva de proceso (ser + participio) y en el 2,8 % a la 3ª persona del plural del indicativo. Asimismo, vimos que, con el recurso preferido (la activa) mantener la perspectiva del recipiente implicaba reformulaciones léxicas importantes en las oraciones españolas con respecto a la oración alemana de referencia que contenía la pasiva de bekommen, pudiéndose calificar las transformaciones subyacentes como modulaciones o expansiones. En oposición a lo anterior, pudimos observar también que cuando se recurría a la pasiva refleja, las reformulaciones en las oraciones españolas podían describirse en la casi todos los casos (90,4 %) como reducciones que implicaban poco coste léxico y a menudo mantenían la perspectiva del beneficiario propia de la pasiva de bekommen. En el mismo trabajo pusimos de manifiesto que la variante más coloquial de la pasiva de bekommen (esto es, la pasiva de kriegen) aparecía en combinación con verbos principales que suponían una acción violenta por parte del agente (no explícito) y que la pasiva de erhalten aparecía en textos pertenecientes a dominios específicos como la economía, las finanzas o la fiscalidad.

El trabajo que se presenta en esta comunicación supone un complemento a los resultados que acabamos de exponer. La actual versión del corpus bilingüe alemán<>español PaGeS (`www.corpuspages.eu`) —con alrededor de 33 millones de palabras procedentes de la traducción directa DE<>ES en su corpus nuclear, v. Doval et al. (2019)— nos permite, por un lado, estudiar el fenómeno de la pasiva de bekommen sobre la base de datos paralelos (ya no paralelizados) y, por otro, incluir el parámetro de la direccionalidad en el estudio. Asimismo, comprobaremos hasta qué punto las preferencias semánticas y pragmáticas de la pasiva de kriegen y de la pasiva de erhalten se comportan en la traducción. Así, continuaremos «ahondando en las diferencias de uso de las formas verbales en la pareja de lenguas implicadas en la traducción» (García Yebra 1997, 347), en este caso el contraste entre las pasivas de bekommen/erhalten/kriegen y las formas verbales españolas correspondientes.

Referencias

Doval, Irene, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas, and Barbara Lübke. 2019. “Corpus PaGeS: A Multifunctional Resource for Language Learning, Translation and Cross-Linguistic

Research." In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, edited by Irene Doval and María Teresa Sánchez Nieto, 103–21. Ámsterdam: Benjamins.

Drosdowski, G. 1995. *Duden: Grammatik der deutschen Gegenwartssprache*. Mannheim etc.: Dudenverlag.

Eggelte, Brigitte. 2020. *Gramática alemana*. Salamanca: Ediciones de la Universidad de Salamanca.

Engel, Ulrich. 1996. *Deutsche Grammatik*. Heidelberg: Groos.

García Yebra, Valentín. 1997. "La voz pasiva francesa y su traducción al español." *Thélème: Revista Complutense de Estudios Franceses* 11: 347–53.

Sánchez-Nieto, María Teresa. 2017. "Wiedergabe der Rezipientenperspektive: Entsprechungen des bekommen-Passivs im Spanischen." *Lebende Sprachen* 62 (1): 187–208. https://doi.org/10.1515/les-2017-0010.

Compilation of DIY parallel corpora in the translation classroom using TAligner

Zuriñe Sanz-Villar CT

Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)

Translation trainees in their later undergraduate years at the University of the Basque Country (UPV/EHU) are familiar with monolingual and parallel corpora. As mentioned by Borja Albi (2019: 23): "[T]housands of ready-made corpora with a variety of purposes, structures and subjects are publically available. Some of them are ready available and can even be searched online using the corpus owner search engine". In the German-Spanish language combination, for instance, the PaGeS corpus is a bidirectional parallel corpus of German and Spanish literary texts that meets the mentioned criteria and is used in our translation teaching setting. However, these students are usually not used to compiling and exploiting their own corpora, although, the reason may not be the lack of IT tools, as "recent developments in software programmes and utilities for corpus building and searching make it possible for any interested person to compile and employ ad-hoc corpora quickly and simply using web con¬tent or their own materials" (Borja Albi, 2019: 23).

The goal of the presentation will be to show how the students created their own corpora, how they used it in combination with a comparable corpus created with AntConc 3.5.9 (Anthony, 2020), and to present the results of students' feedback obtained from a survey conducted at the end of the course. Following the learning corpus use to translate approach (Beeby et al., 2009), students of the very last year of the Degree in Translation and Interpreting at the UPV/EHU compiled their own DIY parallel corpora in a German-into-Spanish specialized translation training class. For that, they used the tool TAligner 3.0 developed within the TRALIMA-ITZULIK research group. This user-friendly tool is free available and its main advantage is that "it allows both corpus alignment and analysis within one tool" (Sanz-Villar and Andaluz-Pinedo, 2021: 142). The corpora consisted of translations done by the students themselves and their classmates. The results discussed refer to the feedback obtained from a survey conducted at the end of the course. Although it was not compulsory to create the corpus, preliminary results indicate that for most of the students the compilation of their own DIY corpus was useful or even very useful.

Keywords: parallel DIY corpora, translation trainees, TAligner 3.0

References

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

Beeby, A.; Rodríguez, P. and Sánchez-Gijón, P. (2009). *Corpus Use and Translating*. Amsterdam: John Benjamins.

Borja Albi, A. (2019): "How corpora can assist legal translation learners: The GENTT TransTools Corpora platform and Sketch Engine". *Quaderns de Filologia: Estudis Lingüístics* XXIV: 21-38. doi: 10.7203/QF.24.16297

Sanz-Villar, Z. and Andaluz-Pinedo, O. (2021): "TAlinger 3.0. A tool to create parallel and multilingual corpora". In Lavid-López, J.; Maíz-Arévalo, C. and Zamorano-Mansilla, J. R. (eds.), *Corpora in Translation*

and Contrastive Research in the Digital Age. Recent advances and explorations. Amsterdam: John Benjamins.

Creating a multilingual parallel corpora: the UV2 web application

Anja Weingart; Georg A. Kaiser; Svenja Schmid

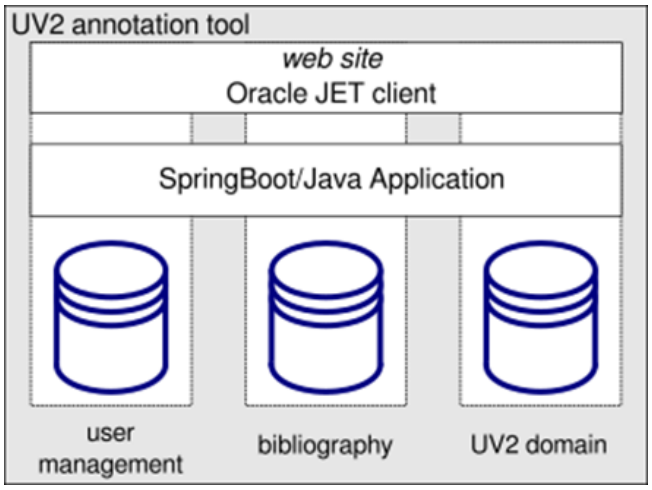
CT

Universität Konstanz

The paper presents the web application *UV2 annotation tool* that allows for the creation and annotation of multilingual parallel corpora. The UV2 tool is developed as part of the DFG project *Uncovering verb-second effects. An interface-based typology* (UV2) (in collaboration with Maia Duguine, (CNRS-IKER UMR5478 Bayonne). Its design is in compliance with the workflow of a non-computational linguist who wants to manage and publish her/his (parallel) data collections. The UV2 tool is a Spring Boot application with a component-based architecture that implements the design pattern Boundary-Control-Entity, Data Transfer Object, and RESTful web services (see Bien 2012, Fielding 2000 Weingart & Cordes 2016). The user interface is a single-page web application built with Oracle JET. The simplified structure is shown (1). The presentation focuses on the UV2 domain, in particular, we will discuss and demonstrate how the alignment of textual elements is managed. Furthermore, we will briefly address how the UV2 tool meets the requirements that derive from the workflow of a non-computational linguist.

(1)

UV2 architecture



We will start with a brief introduction of our corpus of parallel texts that consists among oth-ers of parts of the Bible from the New and Old Testament (Gen-2Kings and Matthew-John) and more than 20 Asterix comic books in the major Romance languages and varieties and in some minor Romance languages (e.g. Asturian, Aragonese, Occitan, Retoromance) as well as in Basque and Latin. Differently to Christodouloupoulos Steedman (2015), our corpus focusses on modern and old Romance languages and varieties. Furthermore, we include the data collected for the project *The structure of wh-utterances and the interpretation of wh-words in Romance (and Germanic) languages* (in collaboration with María Biezma (UMass)). This corpus consists of translations of detective novels and other stories published in '*The Complete Sherlock Holmes*' in the major Romance languages. After preprocessing the texts, each translation is structured in the following tabular format:

(2)

a. MT9:11-SPA ¿Por qué come vuestro Maestro con los publicanos y pecadores?

b. MT9:11-ARA ¿Por qué mincha ro buestro maistro con pecataires y publicáns?

‘Why does your teacher eat with tax collectors and other sinners?’

(3)

a. AST01-5:7b-FRA « Tu reviens bientôt, Astérix ?.. »

b. AST01-5:7b-SPA “¿Vuelves pronto, Asterix…?”

‘Do you come back soon, Asterix?’

(4)

a. SH02-157-SPA ¿por qué no retiró el mismo Jonathan Small aquel tesoro?

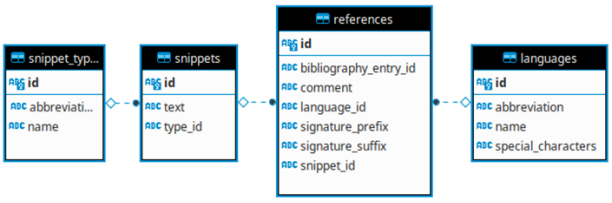
b. SH02-55-CAT ¿per què en Jonathan Small no es va fer amb el tresor pel seu compte?

‘Why did not Jonathan Small get the treasure himself?’

For the Bible corpus we chose the verse as the aligning unit, because sentence boundaries are sometimes unclear, in particular in older translations. In the Asterix and the Sherlock Holmes corpus it is the sentence. This means that there are two types of basic textual units which are called snippets. Each snippet has a unique identifier - a reference - that consists of the attributes, as shown (5).

(5)

entity-relationship model for snippets



The attribute signature_prefix corresponds to the abbreviation of the books/documents, such as MT, AST01, and SH02, and the signature_suffix corresponds to the location of the snippet in the document, for example 9:11 or 5:7b. The signature is the same for each parallel snippet, but only due to the language_id and a bibliography_entry_id, the identifiers become unique. We will demonstrate how the texts are selected and displayed in the user interface only on the basis of a reference endpoint. Starting with such a simple data model for storing textual elements has the advantage that it can be

refined in course of the project and according to the needs of the non-computational linguists who work with the tool.

References

Bien, Adam (2012). *Real World Java EE Patterns*. press.adam-bien.com

Christodouloupoulos, Christos., Steedman, Mark (2015). A massively parallel corpus: the Bible in 100 languages. Lang Resources Evaluation 49, 375–395 (2015).

Fielding, R. Th. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD Diss. University of California, Irvine.

Kaiser, G.A., K. von Heusinger S. Schmid (2019). *Word order variation in Spanish and Italian interrogatives. The role of the subject in 'why'-interrogatives*. In: N. Pomino (ed.), *Proceedings of the IX Nereus International Workshop: "Morphosyntactic and Semantic Aspects of the DP in Romance and Beyond"*. Konstanz: Fachbereich Linguistik, Universität Konstanz (= Arbeitspapier 131), 69–90.

Weingart, A. N. Cordes (2016). *Components of Romance Syntax-O-Matic (CoRS-O-Matic): A free, webbased software application for linguistic data management*. Poster presentation at Going Romance 2016.

Enriching the MUST project: Basque and EN, DE or ES translation pairs

Naroa Zubillaga Gomez CT

Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)

The aim of this presentation is to explain the contribution the research group TRALIMA/ITZULIK will be doing to the MUST, Multilingual Student Translation corpus project, which has become a network of 38 universities, since it first started in 2016. The universities taking part in this project collect translations done by trainee students and upload them to a tailor-made interface, so that the translations can be aligned at the paragraph and sentence level, and errors can be annotated. Among the advantages of this project, already mentioned by Granger and Lefer (2020), I would like to highlight the following: i) the large number of language pairs involved, including minority languages, such as Basque (EU); ii); the interface housing all translations, which is equipped with research as well as teaching functionalities; iii) the language-independent error-annotation system, and iv) the big amount of metadata, not only about the translation task, but most importantly, about the translators, meeting thus a demand for knowledge of the sociocognitive circumstances underlying translational output (Kotze 2020). In Spring 2019 the research group TRALIMA/ITZULIK of the University of the Basque Country joined this project, adding four language pairs, three of them totally new (EN-EU, DE-EU and EU-ES), benefiting from the fact that translation trainees can only train translating into or from Basque at the University of the Basque Country. All the above will be illustrated by commenting on the collection, alignment, and annotation of student translations from EN into EU. To conclude, I will argue that the participation of TRALIMA/ITZULIK in the MUST project has advantages for both sides: our research group will have the opportunity to benefit from the empirical and standardized way of annotating and evaluating students´ translations, while the MUST project will gain with the Basque language brand new language pairs and the corresponding translational as well as metatextual data.

Key words: MUST, Basque, TRALIMA/ITZULIK, trainee students, corpus-based translation studies.

References

Granger, S. & Lefer, M.-A. (2020). "The Multilingual Student Translation corpus: a resource for translation teaching and research". Lang Resources Evaluation 54, 1183–1199.

Kotze, H. (2020). Converging what and how to find out why: An outlook on empirical translation studies. In L. Vandevoorde, J. Daems & B. Defranq (Eds.), *New Empirical Perspectives on Translation and Interpreting* (pp. 333-371). Routledge.

Workshop

Parallel corpora: tools and applications

Johannes Graën

WS

Universität Zürich

The compilation of parallel corpora aims at facilitating the comparison of languages by means of authentic language data. It is thus a necessity to identify nested correspondences on different levels (typically documents, sentences and words). Machine learning methods can be employed to automatically determine the most likely set of those correspondences, referred to as alignments.

Parallel corpora can be employed for several purposes ranging from data-driven research (comparative linguistics, translation studies, typology) to more practical applications (machine translation, bilingual lexicology, language learning).

In this workshop, we will look at existing tools and methods to compile and align parallel corpora, and at some precompiled parallel resources. We will also experiment with tools based on parallel corpora and assess their usefulness for various tasks.

entity-relationship model for snippets