Parallel Corpora for Contrastive and Translation Studies

New resources and applications

EDITED BY

Irene Doval

M. Teresa Sánchez Nieto

Studies in Corpus Linguistics 90

JOHN BENJAMINS PUBLISHING COMPANY

Copyright 2019. John Benjamins Publishing Company. All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

EBSCO Publishing : eBook Collection (EBSCOhost) - printed on 2/10/2023 1:23 AM via
AN: 2090521 ; Doval Reixa, Irene, Sanchez Nieto, Maria Teresa.; Parallel Corpora for Contrastive and Translation Studies : New Resources and Applications
Account: ns335141

Parallel Corpora for Contrastive and Translation Studies

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Studies in Corpus Linguistics (SCL)

ISSN 1388-0373

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/scl

General Editor

Ute Römer
Georgia State University

Founding Editor

Elena Tognini-Bonelli
The Tuscan Word Centre/University of Siena

Advisory Board

Laurence Anthony
Waseda University

Antti Arppe
University of Alberta

Michael Barlow
University of Auckland

Monika Bednarek
University of Sydney

Tony Berber Sardinha
Catholic University of São Paulo

Douglas Biber
Northern Arizona University

Marina Bondi
University of Modena and Reggio Emilia

Jonathan Culpeper
Lancaster University

Sylviane Granger
University of Louvain

Stefan Th. Gries
University of California, Santa Barbara

Susan Hunston
University of Birmingham

Michaela Mahlberg
University of Birmingham

Anna Mauranen
University of Helsinki

Andrea Sand
University of Trier

Benedikt Szmrecsanyi
Catholic University of Leuven

Elena Tognini-Bonelli
The Tuscan Word Centre/University of Siena

Yukio Tono
Tokyo University of Foreign Studies

Martin Warren
The Hong Kong Polytechnic University

Stefanie Wulff
University of Florida

Volume 90

Parallel Corpora for Contrastive and Translation Studies
New resources and applications
Edited by Irene Doval and M. Teresa Sánchez Nieto

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Parallel Corpora for Contrastive and Translation Studies

New resources and applications

Edited by

Irene Doval
University of Santiago de Compostela

M. Teresa Sánchez Nieto
University of Valladolid

John Benjamins Publishing Company
Amsterdam / Philadelphia

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

DOI 10.1075/scl.90

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2018047820 (PRINT) / 2019001810 (E-BOOK)

ISBN 978 90 272 0234 5 (HB)
ISBN 978 90 272 6284 4 (E-BOOK)

© 2019 – John Benjamins B.V.
No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Company · https://benjamins.com

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table of contents

Part I. Parallel corpora: Background and processing

Part II. Parallel corpora: Creation, annotation and access

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Part III. Parallel corpora: Tools and applications

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Acknowledgments

This volume is a selection of studies that were presented at the 1st PaCor 2016 International Conference “Parallel Corpora: Creation and Applications”, held in Santiago de Compostela, with December 1–3, 2016. For the financial support of the conference, thanks are due to the Public Research Agency of the Spanish Ministry of Economy and Competitiveness.

This volume is first and foremost the result of the enthusiastic, committed effort on the part of its contributors, who generously submitted their work to a strict process of peer review evaluation and editorial revision. We thank them for their patience and constructive spirit for the sake of the general coherence of this volume and for their contributions to an excellent work atmosphere.

Very special thanks are owed to Lixun Wang, Carme Colominas Ventura, Clara Pignataro, Victor Darriba Bilbao, Francisco José Ribadas Pena, Manuel Vilares Ferro, María del Mar Sánchez Ramos, Philippe Humblé, Miriam Seghiri Domínguez, Teresa Tomaszkiewicz, Guidor Bilbao Telletxea, Alberto Álvarez Lugrís, Jorge Leiva Rojo, María Calzada Pérez, Inmaculada Serón and Ciprian-Virgil Gerstenberger, who assisted us as external experts in the strict review and evaluation process for the selected papers. We very much appreciate their valuable comments and suggestions, which benefited the final selection of papers to be included here. Their detailed and very constructive report helped us to clarify many points and to improve the structure of the current volume. We are greatly indebted to them all.

Finally, we are deeply grateful to our series editor, Ute Römer, and to Kees Vaes at John Benjamins for their solid, professional support and helpful guidance in shaping this volume. Needless to say all remaining errors are ours.

Irene Doval
M. Teresa Sánchez Nieto

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Parallel corpora in focus

An account of current achievements and challenges

Irene Doval and M. Teresa Sánchez Nieto

In December 2016, the research group SpatiAlEs from the University of Santiago de Compostela (Spain), currently working on the Spanish <> German parallel corpus PaGeS <www.corpuspages.eu>, hosted the conference *Parallel Corpora: Creation and Applications* (PaCor 2016) at the University of Santiago de Compostela.[1] The conference brought together researchers specialized in building parallel corpora with those focused on exploiting such resources for a number of purposes. The event aimed to present the current state of parallel corpus research as a whole, and to highlight those corpus-building projects that work with the Spanish language in particular. The conference's success led to the decision to set up a biennial conference series on parallel corpora. The second edition was held on November 5–7, 2018 at the Complutense University in Madrid, and the third edition will take place in 2020 in Vitoria, at the University of the Basque Country – UPV/EHU.

1. Three decades of parallel corpora in linguistic studies

Since the 1990s, corpus-based and corpus-driven approaches have revolutionized all linguistic disciplines. The first corpus projects, like the Brown Corpus in the early 1960s, were exclusively monolingual, largely English. Mainstream corpora would remain monolingual for another two decades. The absolute pioneer in the field of multilingual corpora was R. Filipovic, who in 1971 finished the Yugoslav Serbo-Croatian–English Contrastive Project, which included half of the texts from the Brown Corpus and their translations into Serbo-Croatian. Although largely

[1] This conference was supported by the State Research Agency (AEI) of Spanish Ministry of Science, Innovation and Universities (FFI2013-42571-P) and by the Department of Economy and Industry of the Galician Government (GI-1954, 2014-PG083). Our thanks go to these institutions for supporting our work.

https://doi.org/10.1075/scl.90.01dov

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

© 2019 John Benjamins Publishing Company

unknown within the linguistic community, this corpus remained as a valuable but isolated initiative without any further progress. The earliest well-known parallel corpus is the Canadian Hansard Corpus, consisting of Canadian Parliament proceedings which had been published in English and French, and whose collection began in the late eighties.

It was the creation of the English-Norwegian Parallel Corpus (ENPC) and its sister project the English-Swedish Parallel Corpus (ESPC) in the early 1990s, which laid the cornerstone for future parallel corpora and had the greatest impact on their further development. In the following years, a number of corpora were compiled according to the ENPC model (Hasselgård 2015: 4). Stig Johansson created the ENPC (Johansson & Hofland 1994) containing original texts in both English and Norwegian with translations into the other language. This project was completed in close cooperation with the Swedish team which compiled the ESPC, using the same design criteria and some of the same original English texts (Aijmer, Altenberg & Johansson 1996: 79ff.).

The rapid development of parallel corpora that followed allowed Lars Borin to state, already in 2002, that "[i]n the last decade or so, parallel corpus linguistics has emerged as a distinct field of research within corpus linguistics, itself a fairly young discipline" (Borin 2002: 1). Since then, workshops and conferences devoted to the creation, annotation and processing of parallel and comparable corpora have multiplied. Along with these developments, a wide variety of parallel corpora have been built for different languages and with different goals.

By far, the most important parallel corpora are the multilingual parallel language resources of the European Union. Steinberger et al. (2014) give a comparative overview of the different multilingual resources provided there. The Europarl Corpus, the most widely used of all of them, contains in its latest release (2012) the proceedings from the European Parliament in 21 languages aligned at the sentence level (Koehn 2005). This corpus was followed up by several other corpora containing documents published by other European Union institutions. These corpora are among the most used text sources for language pair combinations from very distinct languages.

Worthy of special mention is OPUS (Tiedemann 2012), the largest collection of freely available multilingual parallel corpora. It is a growing resource which, along with legislative and administrative texts, mostly from the European Union, includes a substantial amount of newspaper texts and some other smaller collections from various online sources, such as subtitles and technical documentation.

Today, an increasing number of bilingual resources are being released that have been automatically compiled by scraping data from bilingual Internet sites. The most successful among them is Linguee (Linguee GmbH, 2010), a collection of bilingual samples from bilingual websites, mostly from texts pertaining to

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

administrative or commercial domains and covering some 25 languages. These widely used tools can hardly be considered a corpus in the strictest sense, but rather a dictionary enlarged with multilingual online resources, since they lack some of the key features of a corpus, such as a research purpose underlying the collection of texts and a set of criteria for selecting and describing them (Zanettin 2012: 8).

Since their inception, parallel and comparable corpora have been growing progressively in terms of their potential applications, becoming an essential resource in machine translation and multilingual natural language processing. They provide essential training data for statistical translation models, translation memories, or lexical and terminological extraction. On the other hand, within the more strictly linguistic fields of research there are four main areas of application, each with its specific users: basic research in contrastive linguistics and translatology, translation practice, lexicography, and, increasingly, the teaching of foreign languages and translation.

Parallel corpora play a major role in cross-linguistic studies, where "the impact of parallel corpora has been dramatic" (Aijmer 2008: 279). In the 1950s, contrastive studies stimulated attention as a pedagogical tool for predicting difficulties in foreign language teaching and learning. Nevertheless, due to their lack of empirical basis, the analyses were largely limited to the comparison of language systems, and their practical results were disappointing. As a result, interest in contrastive studies began to wane. However, since the 1990s the growing availability of bi- and multilingual corpora led to a great revival of cross-linguistic studies. They provide the empirical basis for identifying similarities and differences between languages and allow for the analysis of linguistic phenomena based on original texts and their translations, or on original parallel texts in the case of bidirectional corpora. This corpus-based approach combines the methodological advantages of computational linguistics and the possibility of contrasting 'parallel' texts in two (and later several) languages (Johansson 2007).

Similarly, with translation studies shifting towards a more empirical paradigm, corpus-based and corpus-driven approaches (Saldanha 2009: 4) have become central to the theory, description, and practice of translation (Zanettin 2012). Their use in the descriptive branch of translation studies resulted from "the convergence between the discovery and justification procedures put forward by Gideon Toury (2012) for the study of translation and the data-driven approach developed by Corpus Linguistics for the study of languages" (Laviosa 2015: 10). Comparable and parallel corpora – the latter, however, traditionally playing a secondary role – have been used as the empirical basis to test (a) the assumed general properties of translated texts (translation universals) or (b) what has been called 'translationese', characterised as "deviance in translated texts induced by the source language" (Johansson & Hofland 1994: 26). In the branch of applied translation studies,

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Marco and van Lawick (2009), show how parallel corpora can help learners develop translator competence.

Even if their primary use is within linguistic and translation research, parallel corpora are also becoming increasingly present in foreign language and translation learning and teaching, providing plenty of translation suggestions through examples of real usage. In the case of intermediate to advanced language learners, where learners are able to decide upon suitable matches for themselves, parallel corpora can even replace a conventional dictionary.

Moreover, a growing number of translator trainers resort to what is known as *CULT methodologies* – Corpus Use and Learning to Translate, see for example Gallego-Hernández (2016), Rodríguez-Inés & Gallego-Hernández (2016) – to either teach their trainees to use corpora productively for real-life translation assignments and documentation tasks, or to teach trainees to observe decision-making processes performed by real translators.

2. Processing and using today's parallel corpora: Some trends

The distinction presented by Borin (2002: 6–7) two decades ago between the two research traditions most interested in parallel corpora –parallel corpora for investigating linguistic phenomena, on the one hand, parallel corpora as a source of data in computational linguistics, on the other – continues to be just as relevant as before. Nevertheless, this fact has not hindered fruitful cross-fertilization between both traditions. Computational linguists optimize parallel corpora for linguistic research while those linguists refining parallel corpora produce high quality parallel material of great use to computational linguists who are able to exploit them thanks to the assistance of machines.

Yet, and quite importantly, both groups of researchers are also regularly and alternatively interested in corpora as *products* (to be drawn upon or researched into), or in the *process* of building or compiling such corpora (whether done manually or through highly automatized procedures). The materials gathered in this book can thus be approached from these two angles: the *process* and the *product* – concepts that will be elaborated upon in the following two sections.

2.1 The process perspective

Traditionally, for those researchers interested in corpora who come from the "linguistic" tradition, the complexity of building a large parallel corpus is no small task, but rather the crux of many a research group, especially if it is going to be used in an extensive project by several researchers and thus accessed and

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

queried via the internet. Setting aside the costs involved (Zanettin 2012: 40–41), the ever-evolving complexity of such technologies makes technical support by computer specialists and engineers almost imperative, and this is the preferred option nowadays if the funding is available. Alternatively, a member of the group has to be technologically savvy enough to envisage those tasks that imply "talking" to the computer. However, current responses to this dilemma by the community of parallel corpus builders within the linguistic tradition are far from homogeneous.

Some groups opt for investing in technical training of one or more of their (linguist) members, who, once back from their "computational journey", are able to disseminate their specialized knowledge at an appropriate level of complexity among fellow linguists (see the approach in Molés-Cases & Oster in this volume). Others seek synergies with computer specialists that help adapt their advances to corpus-related tasks and tools to develop tools that, will eventually render the linguist parallel corpus builder independent from the computer specialist (see Sanz-Villar in this volume). In any case, there seems to be a growing awareness of the importance of networking between groups of linguist parallel corpus builders – such is the stance of the team behind InterCorp, see Čermák (this volume).

The intrinsic design and technological complexity of compiling a large parallel corpus has raised the awareness of the need to reuse already existing parallel corpora and to adapt them to new research endeavors. The idea of "reusing" is taken for granted inside the community of corpus linguists and computer specialists, that is, when the final user of the parallel corpus is the machine – as in the case of the development of machine translation learning systems. Nevertheless, this option is far from evident when the final user of the already existing corpus is *human*, typically a research group made up of linguists.

As Rabadán crucially points out in her chapter, there is a wealth of already built resources that researchers can take advantage of by adapting or upgrading them for new research projects and aims before making the decision to build an entirely new corpus. To this end, presentations of parallel corpora like those in the present volume, or surveys of available parallel text collections and corpora like Mikhailov and Cooper´s (2016: 197–211) might prove useful as a starting point for researchers.

A necessary step towards reusability of corpora in general is to boost the capabilities of a parallel corpus through the addition of annotation layers and/or multimedia extensions. The process of enriching a corpus with theory-informed semantic annotation by linguist corpus builders is far from seamless, as the chapter by Lavid points out: from the selection of semantic categories to be annotated, to the means to ensure intra- and inter-annotator agreement, the whole process needs a careful design if scientific relevance is to be attained. Other examples of successful strategies that have led to harnessing the potential of already existing

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

corpora are provided in Gómez Guinovart's chapter, where the author explains how the potential of the CLUVI corpus for translation research was boosted not only by being semantically annotated but also by being enlarged with multimedia data. Similarly, Ferraresi and Bernardini adopted a multimodal design for their EPTIC corpus from the moment of its conception. These multimedia extensions allow for teaching and researching applications across translation disciplines – not only interpreting but also translation studies, as well as a combination of both.

Yet another different but determined step towards reusability is to design a parallel corpus in such a way that the product is comparable to another parallel corpus. This allows for the replication of studies thanks to these so-called "comparable parallel corpora". This concept was applied by Hareide when designing her Norwegian-Spanish Parallel Corpus and rendering it comparable to PACTRES 1.0.

As most of the corpora presented in this volume have adopted corpus design practices that promote reusability, this volume can and should be interpreted as a call to *reuse* the parallel corpora introduced here – as far as the specific conditions of each project allow for it. The chapters in this volume provide an overview of the main current tendencies in the processing of corpora – understanding here the term *process* in its widest sense, that is involving the compilation, annotation, alignment, indexation and query of parallel corpora.

Alignment is a crucial task in the construction of parallel corpora and an initial step for any exploitation thereof. In most of the corpora presented in this volume, documents are aligned at the sentence level with a wide range of tools. Some of them use several open-source sentence alignment packages (e.g. Hunalign in the case of EPTIC-, LF-Aligner both in PaGeS and PEST, or DéjàVu in the case of COVALT). In some corpus projects, specific alignment tools were created, like TraceAligner for ALEUSKA and InterText for Intercorp, the latter being an open-source tool used also by other projects. Automatic alignment can be manually reviewed – like in the case of the Intercorp core texts or PaGeS – or unsupervised, like in the case of the Intercorp Collections. A great added value that considerably expands the usefulness and applications of parallel corpora is the alignment at the word level. Volk's chapter underlines the importance of world-level alignment not only for refining the query process, but also to improve the accuracy of POS tagging and to serve as basis for further annotation layers, such as dependency parsing and semantic annotation. As for sentence alignment, recent advances point to the development of specific tools that also work in contexts of noisy parallel data, as is the case with BleuAlign (Sennrich and Volk 2011).

Another very important step in the process of setting up a corpus is annotation. For a long time, parallel corpora used to feature little or no annotation at all, consisting of just raw texts. However, linguistic annotation is a means to make the implicit information of a raw corpus explicit, thus allowing the corpus to achieve its full potential in terms of information. An annotated corpus allows for more complex queries that combine character strings and linguistic values in the same query. In this volume, Volk discusses the latest annotation methods for parallel corpora, from standard POS tagging to dependency parsing, as well as the benefits of word alignment techniques for parallel corpus annotation, such as word sense and lemma disambiguation. That leads Volk (this volume) to claim that "the annotation of parallel corpora will be superior in quality to any monolingual annotation".

A common thread throughout the chapters devoted to presentations of parallel corpora is the variety of indexation and query tools drawn upon by corpus developers. These are (i) project-specific developed tools, such as KonText in the case of InterCorp, TextHammer in the case of PEST, TRACE in the case of ALEUSKA, the tools developed to search CLUVI (specifically created to carry out bilingual searches in tagged texts) and the prototype Multilingwis to search multilingual corpora; (ii) CQPweb in the case of COVALT or adaptations of CQPweb as in the case of PACTRES 2.0; (iii) NoSketch Engine as in the case of EPTIC; (iv) query tools which are not specifically designed for corpus management, such as Solr in the case of PaGeS, or which are specific to Natural Language Processing but are not web-based, rather code-based, such as Foma in the case of ALEUSKA.

Furthermore, remarkable efforts are currently being made by parallel corpus developers to broaden the palette of tools for querying parallel corpora beyond standard concordancing in order to allow for analyzing collocations, generating frequency lists and discovering a wide range of statistics, thus boosting the usability of their corpora. A significant example of this trend might be PEST's online query tool, called Texthammer (see also Mikhailov 2016, 2018).

2.2 The product perspective

If we look at the contents of this volume from the "product" perspective, the first thing that might catch the eye of the reader is the diverse nature of the parallel corpora presented in this book. Specifically, three of the chapters deal with bilingual parallel corpora (PACTRES 2.0, PaGeS, CLUVI, and MULTINOT). Four of the chapters are devoted to multilingual parallel corpora (InterCorp, COVALT, EPTIC, PEST, and ALEUSKA). Most of the corpora are of a purely textual nature (PACTRES, PaGeS, InterCorp, ALEUSKA, COVALT, PEST and MULTINOT), and two of them are, as mentioned above, multimodal (EPTIC and CLUVI). All of the corpora are annotated at different levels and most of them include multiple domains, only EPTIC and PEST are specialized corpora dealing with legislative language.

Given that the conference aimed to provide a platform for parallel corpora with Spanish as one of the languages, it should come as no surprise that seven

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

of the parallel corpora presented in this book feature Spanish as one of their languages (PACTRES, PaGeS, InterCorp, ALEUSKA, COVALT, CLUVI, and MULTINOT); in the case of EPTIC, the inclusion of Spanish is planned for the near future (Ferrarresi, personal communication).

Browsing through the aforementioned chapters devoted to corpora allows the reader to perceive some general trends in the products presented therein. The first trend points to the effort to overcome some traditional shortcomings of parallel corpora which are often acknowledged in the literature (see below). The following assessment of the corpus projects presented in this volume may lead to the conclusion that some of these shortcomings have become less of an issue in the course of the last two decades.

One of the traditional drawbacks of parallel corpora is their limited size; see, for example, Tiedemann (2011: 27) and Zanettin (2012: 154). This is quite natural, given that of all the texts that are produced in the world, only a limited amount of them get translated, and into very few languages. As a consequence of this scarce availability, building balanced parallel corpora is a more difficult undertaking than building monolingual corpora, and much more so if such parallel corpora are supposed to be bidirectional. These circumstances may be changing, however, with the appearance of new parallel corpus projects clearly investing on size, like InterCorp, PaGeS or PACTRES 2.0.

As for domain specificity, several of the corpora discussed in this book are proof that this limitation is gradually being overcome thanks to the growing availability of electronic texts, which for years had been restricted to legal and legislative discourse, or to material from newspapers or newswire-style texts. The current availability of larger amounts of electronic texts allows corpus builders to draw from a broader range of textual (including multimodal) materials, notably works of fiction along with some non-fiction (see Intercorp, PaGeS, COVALT, PACTRES 2.0, MULTINOT), interpreted discourse (EPTIC), or movies and TV subtitles (CLUVI).

Thanks to this greater availability, some multidomain corpora like MULTINOT and PACTRES 2.0, or InterCorp either specifically achieve or at least aim for a certain level of balance in their choice of genres and translation directions, in accordance with the theories of principled corpus building – e.g. Biber (1993). Yet the nature of other parallel corpus projects, like PEST in this volume, excludes such balance from its aims, as it runs counter to the principle of maximizing size, to which some projects clearly attach more importance. Reactions to this dilemma (principled corpus building vs. opportunistic corpus building) are especially apparent throughout the chapters in the second part of the book.

Nowadays, parallel corpora are not limited to just a few language pairs anymore. Projects like InterCorp boast 40 different languages and allow for parallelizing

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

texts in under-resourced language pairs, for which only very scarce parallel evidence exists (e.g. Vietnamese-Greek). CLUVI, for its part, includes Galician, one of Spain's co-official, minority languages; the same applies to ALEUSKA with Basque, and to COVALT with Catalan.

Finally, as stated above, in terms of annotation, parallel corpora do no longer lag behind monolingual corpora. All of the corpora presented in this volume are annotated at one level or another, with a wide variety of tools, either specifically tailored for the purpose of the project or generally available. What is particularly remarkable is the consistent annotation of some of the corpora presented here, mostly the result of manual labour. Besides POS tagging, several parallel corpus projects aim to provide their textual base with additional layers of annotation, exemplified in this book by CLUVI and MULTINOT. A part of the CLUVI corpus is semantically annotated (VEIGA). The tagging is based on the synsets already established in WordNet. In MULTINOT, various discourse phenomena are annotated by means of a tagset specifically developed for the aims of the project.

Technological advances, the development of corpus linguistics and applied linguistics, and the existence of long-term, funded research projects no doubt form the rich soil on which parallel corpora have thrived as products throughout the last two decades – be it as research aids or as objects of research themselves. Nevertheless, there are still plenty of research contexts where such conditions are not necessarily met. Thus, it is sometimes necessary to resort to comparable corpora, which are much easier to compile since compilers can exploit already existing monolingual corpora to that end. Notably, comparable corpora allow for the user to contrast languages or language varieties for which little to no translation evidence is readily accessible, such as authentic spoken language, or under-resourced and minority languages.

To round off this brief survey from the "product" perspective in this volume, we will now outline the questions raised by the chapters that deal with corpora as specific products with applications for different research areas. In the expansive field of computational linguistics, parallel corpora serve an ever-increasing range of purposes, exemplified in this volume by the contributions by Garcia et al., Gamallo, and Ghoshal and Rao. Garcia et al. present a new method to automatically extract bilingual collocation equivalents from parallel corpora, which in turn are very useful resources for a variety of tasks such as machine translation or the compilation of teaching materials for foreign languages. A great achievement for minority languages, given their usually limited resources for building parallel corpora, is Gamallo's proposal to exploit comparable corpora for the construction of bilingual dictionaries. Finally, Ghoshal and Rao use parallel corpora from two varieties of French (standard French and text message French) for the text normalization of abbreviations and shorthand forms used in text messages.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

In the areas of translator training and language teaching and learning, parallel corpora have, until recently, had rather limited use as "products" to be used in the classroom. There are two general beliefs about corpora that limit the use of parallel corpora in language learning contexts – leaving aside the issues of a possible lack of electricity and access to the Internet that can eventually become real barriers for their use. The first belief is that the kind of language found in corpora is too difficult for learners to understand or to make use of. The second belief is that even if the language were of an appropriate level, the tools for searching parallel corpora are not learner-friendly, as they are designed – so the argument goes – with researchers and not language learners in mind. Moreover, parallel corpora specifically face issues such as a lack of awareness even among generally corpora-aware instructors, and a lack of accessible tools for searching parallel corpora. In this respect, several corpora (PACTRES, PaGeS, COVALT) aim to bridge the gap "between corpus research and pedagogical practice" (Römer 2009: 95) and they focus on building useful and user-friendly resources for language learning and translator training.

Regarding the applications of parallel corpora within translation studies, Marco aligns himself with those authors who demand more space for parallel corpora in descriptive translation studies, and demonstrates in two small-scale pieces of research what this collaboration might look like. Hareide, on her part, in the face of the growing body of research within corpus-based descriptive translation studies, points to the methodological problems that may arise if the same translation-related hypothesis is to be tested in two different parallel corpora. Finally, Sanjurjo-González and Izquierdo acknowledge that parallel corpora play a decisive role in *truly joint* functional contrastive-translation studies, that is those that, starting from a specific form in language A, seek to find discursive equivalents in language B that originated from translations into that language – and demonstrate it with a case study performed using PACTRES 2.0.

3. Structure of this volume and presentation of contributions

The first part of the book is comprised of chapters devoted to the new roles that parallel corpora can and should play within translation studies and contrastive linguistics – be it on their own or in conjunction with comparable corpora –, specific issues of usability and usefulness of parallel corpora, and the latest developments in parallel corpus alignment at the word level.

After a systematic overview of the uses that parallel corpora have had so far in translation studies (sense-disambiguation, the quest for translation universals, help for practitioners and translator training), J. Marco illustrates, with two case studies performed using the COVALT corpus, two main descriptive uses of

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

parallel corpora in this area: as a main source of data to show translation choices or relationships, and as a secondary source of data to complement the findings from research in comparable corpora.

R. Rabadán calls for more sustainable practices in working with parallel corpora, encouraging researchers to at least assess the possibility of drawing from already existing parallel corpora; whether to reuse, reprocess or enlarge them for their own projects. She helps to assess the usefulness of such resources by means of a checklist to identify design-related, linguistic, and technical issues of existing parallel corpora. Furthermore, she calls on the corpus building community to collaborate in an effort to construct new, more reusable corpora, creating validated and standardized resources with replicable corpus building protocols.

Following Rabadan's approach, L. Hareide's chapter focuses on the replicability of corpus-based studies by using "comparable parallel corpora", an innovative concept that allows for the testing of the same hypothesis on two different language pairs, thereby ensuring the basic scientific principle of replicability. The author outlines how her corpus (NSPC) was compiled to be comparable to the English-Spanish P-ACTRES corpus. Both of them are used to successfully test one of the universals of translation hypothesis, namely, the gravitational pull hypothesis.

M. Volk's chapter deals with word alignment in and information retrieval from parallel corpora. Firstly, the author briefly discusses the techniques used by his team in their parallel corpus projects at the University of Zurich: advanced techniques for word-level annotation on the one hand, and word-level annotation techniques for specific languages on the other. The final part of the chapter focuses on information retrieval from word-aligned parallel corpora and the possibilities that word alignment opens up for translation and contrastive studies.

The second part of the book consists of presentations of a number of cutting-edge parallel corpus projects currently underway in Europe. In addition, most chapters in this part present case studies developed on the basis of the corpora in question. In some of these chapters, special attention is being paid to specific technical issues of corpus building as well.

P. Čermák presents the current version (9th Release) of InterCorp, with 40 languages one of the most extensive existing parallel corpora, Czech being the pivot language. This searchable online corpus is comprised of about a hundred billion words. The author offers a general description of the corpus' design and of the tools used to create and annotate it, all of them open-source. Čermák also introduces Intercorp's use and applications. Special attention is given to the sub-corpus of Spanish texts, the second largest in the corpus.

I. Doval, S. Fernández Lanza, T. Jiménez, E. Liste Lamas and B. Lübke present PaGeS, the result of an ongoing project that has resulted in the most extensive searchable online parallel corpus with translations between German and Spanish.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

The project's main goal is to provide a broad base of empirical data without the shortcomings of already existing resources, most of them restricted to legal and administrative language. The project pays special attention to the quality of the originals. The corpus architecture, annotation, pre-processing, alignment, and indexing process are all carefully detailed. PaGeS architecture renders it suitable not only for linguistic and contrastive research, but also for empirical studies of translations along with language teaching.

A. Ferraresi and S. Bernardini present EPTIC, a freely available corpus made up of EU Parliament Proceedings, which is comprised of the transcripts of original speeches, the transcripts of their interpretations, and the official written versions with their official translations, both published online as so-called "verbatim reports". The different text-to-text and text-to-video alignments result in 14 separate subcorpora. EPTIC features three languages (EN, FR. IT), two communication modes (written/spoken), and two forms of mediation (translated/interpreted), and provides access to time-synced videos.

X. Guinovart's chapter provides a detailed description of the CLUVI corpus, the most extensive parallel corpus of the Galician language, containing over 44 million words from a variety of domains with 12 different language combinations. The corpus can be freely searched online. This chapter walks the reader through the complex procedures required for building multilingual corpora with an added value. There are two main foci: the methodology developed for expanding a subcorpus composed of subtitles with multimedia data and the different resources used to build SensoGal, a semantically annotated subcorpus using WordNet synset categories.

In her chapter dedicated to the MULTINOT corpus (EN<>ES), J. Lavid underscores the demand for corpora annotated with discursive-semantic features such as thematization, modality and evidentiality, metadiscourse, rhetorical realizations, appraisal, and opinion. Lavid provides insights into the challenges posed by the manual annotation of MULTINOT that were tackled by means of the project's 6-step scientific annotation procedure. The fact that semantic categories had to be annotated for two different languages only demonstrates the detail involved in the approach.

M. Mikhailov, M. Santalahti and J. Souma describe PEST (Parallel Electronic Corpus of State Treaties), a full text, diachronic, aligned, morpho-syntactically annotated parallel corpus specifically built to study state treaties as a juristic genre closely bound to translation. PEST features state treaties in its Finnish<>Russian, Finnish<>Swedish, and Swedish<>Russian subsections, as well international treaties in an EN-RU-FI-SV section. The authors point to the methodological issues raised by their material – the natural imbalance and skewedness of their corpus – and how this issue may be best compensated by investing in size, annotation

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

with a wide array of metadata, and flexible tools that allow the user to search by subcorpora.

Researchers or teams assessing the resources available to build their own corpora might benefit from T. Molés-Cases and U. Oster's step-by-step guide to compiling and indexing, and creating access to a parallel corpus with IMS Open Corpus Workbench tools and the related web-based graphic user interface of CQPweb, taking the EN > ES section of the COVALT corpus as the basis for their demonstration.

In the context of the research developed by the ACTRES group in functionally oriented EN/ES contrastive and translation studies, H. Sanjurjo-González and M. Izquierdo explain how the EN-ES parallel corpus PACTRES 1.0 was enlarged and technically improved, leading to PACTRES 2.0, detailing compilation criteria and discussing issues of representativity and balance between the ES > EN and the EN > ES subcorpora. Their case study, which rounds off the chapter, demonstrates how PACTRES 2.0 can be exploited as both a comparable and a translational corpus.

Z. Sanz-Villar's translation-driven approach to corpus building resulted in ALEUSKA, a highly original trilingual corpus of German literary texts and their direct and indirect translations into Basque (the latter through the medium of Spanish). In a flexible approach to corpus building, the author exploited a shallowly annotated version of ALEUSKA, and subsequently reused other tools to further annotate her textual base (POS tagging with IXA pipe tools) and to exploit such annotation (Forma programme).

The third and final section of the book discusses specific tools and applications of parallel corpora. P. Gamallo's chapter focuses on what can be done when parallel corpora are not available. He poses two alternative strategies for generating bilingual lexicons from exclusively comparable corpora: either by transitivity via intermediary dictionaries or by means of string similarity for the extraction of bilingual cognates. In both cases, the results are filtered using distributional semantics. Several experiments show that the performance of both strategies is similar to other approaches relying on parallel data.

M. Garcia, M. García-Salido and M. Alonso-Ramos' chapter centers on the use of parallel corpora not only to extract bilingual collocation equivalents, but also to create a bilingual distributional semantic model. This model is used to extract bilingual collocation equivalents from corpora through dependency parsing to identify syntactically related words, while also employing statistical measures for ranking the extracted candidates in order to obtain monolingual collocation candidates. They then train a bilingual model of distributional semantics which is finally applied to the monolingual collocations to automatically identify bilingual equivalents of both the base and the collocate.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

In the final chapter Goshal and Rao use parallel corpora for the very justified practical purposes of normalizing the colloquial and abbreviated text of SMS messages. The authors propose an approach for Belgian French shorthand forms, applying two strategies: the character-based SMT (Moses) and word embeddings (Multivec) that encode contextual information. The authors argue for the combination of these two approaches in order to increase both recall and the precision in the the normalization of shorthand forms in SMS texts. They also highlight the flexibility of the method since it can be easily adapted to text normalization in other domains.

We hope that the contributions to this book will draw the attention of readers as diverse as scholars and undergraduates along with MA and PhD students in the fields of contrastive linguistics, lexicography, language teaching and learning, translation and interpreting studies, both descriptive and applied.

References

Aijmer, Karin, Altenberg, Bengt & Johansson, Mats (eds). 1996. Text-based contrastive studies in English. Presentation of a project. *Lund Studies in English* 88, 73–86.

Aijmer, Karin. 2008. Parallel and comparable corpora. *Corpus linguistics. An International Handbook* Lüdeling, Anke & Kytö, Merja (eds), 275–292. Berlin: Walter de Gruyter. https://doi.org/https://doi.org/10.1515/9783110211429

Aijmer, Karin & Altenberg, Bengt (eds). 2013. *Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson* [Studies in Corpus Linguistics 54]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.54.02intro

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257. https://doi.org/10.1093/llc/8.4.243

Borin, Lars. 2002 "…and never the twain shall meet?". In *Parallel Corpora, Parallel Worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*, Lars Borin (ed.), 1–43. Amsterdam: Rodopi.

Gallego-Hernández, Daniel (ed). 2016. *New Insights into Corpora and Translation*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Hasselgård, Hilde. 2015. Parallel corpora and contrastive studies. In *Proceedings of the international symposium on Using Corpora in Contrastive and Translation Studies 2010 Conference (UCCTS2010)*. <http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/papers/Hasselgard.pdf> (26 June 2018).

Johansson, Stig & Hofland, Knut. 1994. Towards an English-Norwegian parallel corpus. In *Creating and Using English Language Corpora*, Udo Fries, Gunnel Tottie & Peter Schneider (eds), 25–37. Amsterdam: Rodopi.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies* [Studies in Corpus Linguistics 26]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.26

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit* X(5), 79–86. Phuket.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Laviosa, Sara. 2015. Corpora and holistic cultural translation. In *Corpus-based Translation and Interpreting Studies: From Description to Application / Estudios traductológicos basados en corpus: De la descripción a la aplicación*, María Teresa Sánchez-Nieto (ed.), 31–51. Berlin: Frank & Timme.

Linguee GmbH. 2010. *Linguee.com. The Web as a Dictionary*. <http://www.linguee.com> (20 May 2018).

Marco, Josep & van Lawick, Heike. 2009. Using corpora and retrieval software as a source of materials for the translation classroom. In *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate* [Benjamins Translation Library 82], Allison Beeby, Patricia Rodríguez-Inés & Pilar Sánchez-Gijón (eds), 9–28. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.82.03mar

Mikhailov, Mikhail. 2016. TextHammer corpus tool: Not only concordancing. Presentation held at the Next Generation Translation Tools Conference, University of Swansea (15th June 2016). <http://language-research-centre.swan.ac.uk/wp-content/uploads/2016/08/mikhailov.pdf> (11 May 2018).

Mikhailov, Mikhail. 2018. TextHammer, Ver. 1.5 User Manual. <https://puolukka.uta.fi/texthammer/texthammer_1_5_en.pdf> (11 May 2018).

Mikhailov, Mikhail & Cooper, Robert. 2016. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. London: Routledge. https://doi.org/10.4324/9781315624570

Römer, Ute. 2009. Corpus research and practice: What help do teachers need and what can we offer? In *Corpora and Language Teaching* [Studies in Corpus Linguistics 33], Karin Aijmer (ed.), 83–98. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.33.09rom

Rodríguez-Inés, Patricia & Gallego-Hernández, Daniel. 2016. Corpus Use and Learning to Translate, almost 20 years on. *Cadernos de Tradução* 36(1): 09–13. https://doi.org/https://doi.org/10.5007/2175-7968.2016v36nesp1p9

Saldanha, Gabriela. 2009. Principles of corpus linguistics and their application to translation studies research. *Tradumàtica: Traducció i tecnologies de l'informació i la comunicació 7*. <http://www.fti.uab.cat/tradumatica/revista/num7/articles/01/central.htm> (11 May 2018).

Sennrich, Rico & Volk, Martin. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of The 18th International Nordic Conference of Computational Linguistics (Nodalida)*, 175–182. Riga.

Steinberger, Ralf et al. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation* 48(4): 679–707. https://doi.org/10.1007/s10579-014-9277-0

Tiedemann, Jörg. 2011. *Bitext Alignment*. San Rafael CA: Morgan & Claypool Publishers.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2214–2218. <http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf> (20 May 2018).

Toury, Gideon. 2012. *Descriptive Translation Studies and beyond: revised edition*. Amsterdam: Benjamins.

Xiao, Richard. 2010. *Using Corpora in Contrastive and Translation Studies*. Newcastle upon Tyne: Cambridge Scholars.

Zanettin, Federico. 2012. *Translation-driven Corpora*. London: Routledge.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART I

Parallel corpora

Background and processing

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Comparable parallel corpora

A critical review of current practices in corpus-based translation studies

Lidun Hareide
Møreforsking Molde, Norway

Are papers presented in corpus-based translation studies truly scientific? These are normally done on only one language pair, often on purpose-made parallel corpora, and can normally not be replicated. Therefore their value is limited in a strictly scientific sense. The use of comparable parallel corpora allows both for the replication of studies, and the testing of complex hypotheses like Halverson's Gravitational Pull hypothesis. This chapter defines and discusses the concept of comparable parallel corpora, and exemplifies their value by illustrating their use. The chapter also presents hopes for the future, as new groundbreaking technology that will allow the linguist to create her own parallel corpora without the aid of computer scientists is currently being launched at the University of León in Spain.

Keywords: comparable parallel corpora, the Gravitational Pull Hypothesis, unique items

1. Introduction

As Geoffrey Leech (1992: 112) so aptly stated, "a significant advantage of the corpus linguistic methodology is that it allows for the analyst to approach the study of language from the context of the scientific method". Just as the incorporation of electronic corpora revolutionized the field of grammar, it also caused what has come to be known as the "empirical turn" in translation studies (Snell-Hornby 2006: 115). Chesterman (2004: 46) pointed out that "Corpus-based research into translation universals has been one of the most important methodological advances in translation studies during the past decade or so, in that it has encouraged researchers to adopt standard scientific methods of hypothesis generation and testing". Since its inception, Corpus Based Translation Studies (CBTS) has been

https://doi.org/10.1075/scl.90.02har

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

© 2019 John Benjamins Publishing Company

one of the fastest growing subfields of translation studies (Ji et al. 2017: 4). One must question, however, just how scientific the empirical turn has turned out to be, as even today most parallel corpora used for translation studies are compiled to solve specific research questions, and are not comparable to other parallel corpora. This often means that the research done on these corpora cannot be replicated on other language pairs, and therefore, in a strictly scientific sense, the value of these studies is limited.

Just as worrisome are studies where broad and general conclusions are drawn from findings from a very limited (and often very specific) corpus. It is time to take a step aside to critically review our own practices, remind ourselves of the basic principles of corpus linguistics and to propose some improvements.

This chapter discusses the concept of comparable parallel corpora (Hareide 2012, 2014, 2017). Comparable parallel corpora allow both for the replication of studies, and for the testing of complex hypotheses, such as the gravitational pull hypothesis (Halverson 2003, 2007, 2010). In this chapter, I will first define the concept of comparable parallel corpora, then give a brief overview of the subfield of translation universals of which Halverson's hypothesis forms an integral part, and briefly outline how the Norwegian Spanish Parallel Corpus (NSPC) (Hareide & Hofland 2012) was compiled in order to be comparable to version 1 of the English – Spanish P-ACTRES corpus developed at the university of León in Spain (Rabadán 2005, 2007, 2005–2008; Izquierdo, Hofland & Reigem 2008: 31; Rabadán, Labrador & Ramón 2009) and to the CREA reference corpus,[1] before describing how parallel comparable corpora were used to successfully test the gravitational pull hypothesis (Hareide 2014, 2017b). Some suggestions for the future will conclude this chapter.

2. Comparable parallel corpora

Altenberg and Granger (2002: 7–8) distinguish between parallel corpora and comparable corpora in the following way: A bilingual parallel corpus is defined as "original texts in one language and their translation into one or several other languages" whereas comparable corpora are defined as "original texts in each language, matched as far as possible in terms of text type, subject matter and communicative function". Building on these definitions, *comparable parallel corpora* are defined as:

[1] Real Academia Española: Banco de datos (CREA) [en linea]. *Corpus de referencial del español actual* <http://www.rae.es>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

> Two or more corpora containing original text in one language and the published translations of these texts into at least one other language, that are matched as far as possible in terms of sampling frame, that is text-type, subject matter and communicative function as well as time-frame and the language variety sampled.
> (Hareide 2017: 189, based on Hareide 2014: 209 and Hareide & Hofland: 2012)

A shorter version of this definition would be simply:

"Two or more parallel/translation-corpora that have the same sampling frame". In his influential paper *Representativeness in Corpus Design,* Biber defined a sampling frame as: "an operational definition of the population, an itemized listing of population members from which a representative sample can be chosen" (Biber 1993: 244). I will return to the specifics of sampling frames in Section 5.

Comparable parallel corpora must be made up of data from two separate language pairs in order to allow for testing of the same hypothesis on two different language pairs, thereby ensuring the basic scientific principle of replicability. When used for testing Halverson's gravitational pull hypothesis, the two language pairs must have the same target-language but different source languages. I will return to this hypothesis and the measures taken to devise a method for its testing later in this chapter, but first a brief overview of the developments leading up to its formulation is needed, starting where it all started, with Baker's (1993) suggestion that corpus linguistics could be a serviceable method for testing hypotheses in Translation Studies.

3. Universals of translation

Baker's influential paper opened up a new world of empirical quantitative research for translation scholars (Chesterman 2010: 40). With the aid of electronic corpora, large quantities of translated text could be searched for traits that are specific to translations, and a very productive field of study, which came to be known as Corpus Based Translation Studies (CBTS) saw the light of day. Although the field of study has broadened considerably in the decades following this publication, Baker initially suggested using large corpora to study the linguistic nature of translations, either by contrasting them to their source texts or to untranslated target-language texts (Mauranen & Kujamäki 2004: 1). In her groundbreaking paper, Baker identified six "features which typically occur in translated text rather than in original utterances and which are not the result of interference from specific linguistic systems" (Baker 1993: 243). These features were hypotheses that other scholars had presented based on small-scale, manually conducted comparative studies, and Baker's contribution was to propose the use of corpus-based methods

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

to empirically investigate whether these represented translation universals or not. Some of the features she listed are: a tendency towards explicitation (spelling out or adding information) (Blum-Kulka 1986; Toury 1991), a tendency towards disambiguation and simplification (Blum-Kulka & Levenston 1983; Vanderauwera 1985), a strong preference for conventional "grammaticality" (Schlesinger 1991; Vanderauwera 1985), avoidance of repetition (Shlesinger 1991; Toury 1991), specific relations between specific source and target languages and a tendency to exaggerate features of the target language (Baker 1993: 244–247).

These six hypotheses came over the years to be collectively known as the translation universals hypothesis. Although it has been widely criticized, the search for universal features has brought important new momentum to the field of translation studies (Chesterman 2004: 46). One of these features, the hypothesis that one can observe "a general tendency to exaggerate features of the target language" (Baker 1993: 244, based on Toury (1980), Vanderauwera (1985), and Shlesinger (1991)) has been considered one of the most controversial and most interesting features from a research perspective. It is often referred to as the overrepresentation hypothesis, since Vanderauwera suggested that translations "over-represent features of their host environment in order to make up for the fact that they were not meant to function in that environment" (Baker 1993: 245). Empirical research by Halverson (2007), among others, on the Norwegian structures that give rise to the English progressive in translations supports this hypothesis.

Baker's Overrepresentation Hypothesis was opposed by Tirkkonen-Condit (2002, 2004), who named target-language specific features "unique items" and proposed the unique items hypothesis. She claimed that these language structures are in fact under-represented in translations because there are no structures in the source language that will trigger their use (Tirkkonen-Condit 2004: 177). Tirkkonen-Condit claimed that translations tend to contain fewer of these unique items, and their frequency in a text can determine whether the reader believes the text to be an original or a translation (Tirkkonen-Condit 2004: 178). Tirkkonen-Condit conducted her study of Finnish verbs of sufficiency and Finnish clitic particles on the Corpus of Translated Finnish (CTF), and her hypothesis is supported by empirical research by, among others, Kujamäki (2004), Eskola (2004), Rabadán, Labrador & Ramón (2009), Vilinsky (2012) and Capelle (2012).

This state of affairs left the field of Translation Studies with two conflicting hypotheses, both supported by extensive research. Halverson's (2003, 2007, 2009, 2010) gravitational pull hypothesis, however, aimed to predict and explain how these two outcomes of the translation process proposed by Baker and Tirkkonen-Condit can be expected in different situations. Halverson stated that Baker's and Tirkkonen-Condit's hypotheses are not specific enough in their predictions; and it is neither clear what circumstances lead to translational overrepresentation

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

or underrepresentation (Halverson 2010: 364) nor what characteristics in the translation process and in the languages in question make different language pairs relate to each other differently. From the standpoint of cognitive grammar, she suggested that both overrepresentation and underrepresentation of particular target-language items is possible. However, the likelihood of a particular translated outcome (e.g., overrepresentation or underrepresentation) will depend on the specific structure of the bilingual semantic network activated in any given instance, and specific configurations will predict specific translational outcomes (Halverson 2010: 352).

In Figure 1 the way the potential sources of translation effects operate and interrelate is visualized, and I will explain the hypothesis by adopting Halverson's (2007: 176) operationalization of the notion of Gravitational Pull with reference to grammatical structures, in this case the Spanish gerund, in translations from Norwegian and English. Halverson (2010: 356) identified three factors that can lead to underrepresentation or overrepresentation. The first factor, patterns of prototypicality, reflects the way the target language is used. If a particular element is frequently used or prototypical in the target language, the hypothesis predicts that this structure will exert gravitational pull, and will therefore be overrepresented.

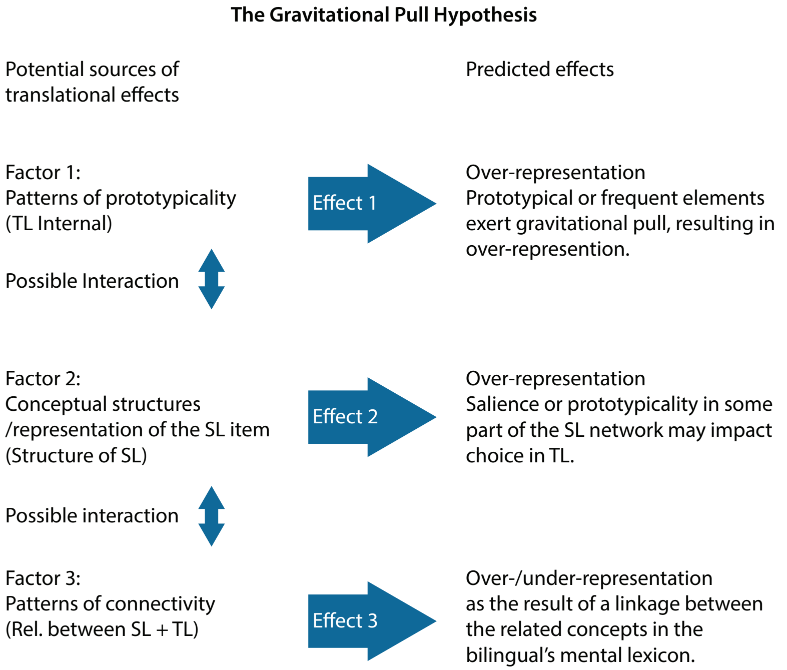

The Gravitational Pull Hypothesis
Potential sources of translational effects
Predicted effects
Factor 1: Patterns of prototypicality (TL Internal)
Effect 1
Over-representation Prototypical or frequent elements exert gravitational pull, resulting in over-represention.
Possible Interaction
Factor 2: Conceptual structures /representation of the SL item (Structure of SL)
Effect 2
Over-representation Salience or prototypicality in some part of the SL network may impact choice in TL.
Possible interaction
Factor 3: Patterns of connectivity (Rel. between SL + TL)
Effect 3
Over-/under-representation as the result of a linkage between the related concepts in the bilingual's mental lexicon.

Figure 1. The gravitational pull hypothesis based on my understanding of Halverson (2010). First published in Hareide (2014)

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

The Spanish gerund is very frequent (4,348 gerunds per million words in the subset of the CREA reference corpus used in this study), and according to the Spanish reference grammar the *estar* + gerund structure is both very frequent and also perceived as the prototypical representation of the concept of the Spanish gerund (RAE 2009: 2186). The gravitational pull hypothesis therefore predicts that both the Spanish gerund and *estar* + gerund structures will be overrepresented in translations both from Norwegian and English in comparison to in text originally written in Spanish (Hareide 2017: 194).

The second factor is called "conceptual structures/representation of the source-language item" and reflects whether the source language items are represented as prototypical or salient in the translator's mind. As outlined in Hareide (2017b), the *estar* + gerund structure corresponds structurally to the English progressive. In several studies of second language acquisition and languages in contact, the English *-ing* form and its subcategory, the English progressive, are reported to be salient structures (Goldschneider & DeKeyser 2005: 310; Gass 2008: 36, 145; Sanchez 2006: 310), and Halverson herself postulates the *be* + V-ing construction as "a highly salient prototype for the expression of proximity to the speech event" (Halverson 2007: 183). Being both salient and having prototypical status, the English progressive may influence the translator to use the Spanish gerund in translations from English and will exert a pull towards overrepresentation. In Norwegian, however, comparative grammars and the empirical evidence presented in Hareide (2017a) point to no direct grammatical counterpart of the Spanish gerund in Norwegian, and consequently no salient structures that correspond to the Spanish gerund (see Hareide 2017a Section 2.2), and factor 2 does therefore not contribute to the overrepresentation of the Spanish gerund and the *estar* + gerund in translations from Norwegian.

It is the third factor that constitutes a test of the mutually exclusive overrepresentation hypothesis (Baker 1993, 1996) versus the unique items hypothesis (Tirkkonen-Condit 2002, 2004). Here the degree of linkage ("that is, the "distance" or "degree of overlap" (Halverson 2010: 356)) between related concepts in the two languages in the bilingual's mental lexicon may result in a pattern of either overrepresentation or underrepresentation, depending on "the presence or absence of direct links between certain network configurations and lexical items in the bilingual's languages" (Halverson 2010: 355–356).

To test for this factor, a language pair with a unique item is required. Remember, no indication of any direct grammatical counterpart was found neither in comparative grammars nor in the empirical evidence presented in Hareide (2017a) and therefore testing using this structure on the language pair Norwegian–Spanish is ideal. We would expect underrepresentation in accordance with the unique items hypothesis and factor 3 of the gravitational pull hypothesis to be the result.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

For the English–Spanish language pair the situation is very different as contrastive grammars and studies point to some grammatical "overlaps" between English grammatical resources and the Spanish gerund. These overlaps are of two kinds: either a formal overlap where the same structure exists in both languages, or a functional overlap where different grammatical structures in the two languages perform the same function. The Spanish *estar* + gerund and the English progressive *be* + V-*ing* are both progressive structures and as such have a considerable formal overlap, which is explored in Hareide 2017b. The overlap between these two grammatical structures in the two languages may be perceived in the bilingual person's mental lexicon as a linkage – or even a direct link – between these two structures. In addition, the Spanish gerund has a functional overlap with English non-finite adverbial phrases, as it is believed to perform many of the same functions as these (see Hareide 2017b 2.2.3), which also may be perceived as a linkage in the bilingual person's mental lexicon. These mental overlaps will, according to the gravitational pull hypothesis, result in overrepresentation of both the Spanish gerund and its subcategory the *estar* + *gerund* in accordance with Baker's hypothesis, because the overlapping forms and direct links constitute triggers for the use of the Spanish gerund.

However, the overlap between the English and Spanish resources is only partial, and forms that do not overlap will be expressed by other grammatical or lexical structures. In cases where neither overlaps nor direct links are perceived, which is the case in translations from Norwegian, there is no source-language item to trigger the item in question, and this factor will not be activated. The result may be underrepresentation in accordance with Tirkkonen-Condit's hypothesis – for a full description of the hypothesis, see Halverson (2010) or Hareide (2017b). In order to be fully tested, Halverson's gravitational pull hypothesis requires two separate data sets from two language pairs or parallel corpora, that have the same target language, but different source languages. These two source languages must relate to each other differently, that is the linguistic item to be tested must constitute a unique item in one of the language pairs, but not in the other. In order to compare these data sets, the corpora they are drawn from need to be comparable, hence the term "Comparable parallel corpora".

4. What makes parallel corpora comparable, and does size matter?

As may be remembered from the definition in Section 2, what makes parallel corpora, such as the P-ACTRES and the NSPC, comparable, is their sampling frame; that is the predefined criteria for the selection of texts to be included in the corpus. Here the most important question to ask is: "what universe of text is this

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

corpus intended to represent? (Biber 1993: 243). According to Biber, there are at least two aspects to consider when defining the target population (i.e. the universe the texts in the corpus will be sampled from). First, the boundaries of the target population – ”what texts are included and excluded from the population” (Biber 1993: 243) and second, “what text categories are included in the target population and what are their definition” (Biber 1993: 243). Biber warns that:

> In designing text corpora these concerns are often not given sufficient attention, and samples are collected without a prior definition of the target population. As a result, there is no possible way to evaluate the adequacy or representativeness of such a corpus (because there is no well-defined conception of what the sample is intended to represent) (Biber 1993: 243)

When defining the parameters for the target population (the sampling frame), compilers have some liberty but some common categories include the following:

- time frame of publication
- the language variety sampled
- genres/text types represented
- specialized vs. non-specialized language
- written vs. oral (transcribed) texts
- original vs. translated language
- size

The importance of a well-defined sampling frame cannot be stressed enough, and it is worth remembering that as early as in 1998 Tymockzo (1998: 6) questioned the quality, composition, and representativeness of corpora, and to what extent results from these studies can be generalized to other language pairs and replicated (1998: 3–8). Similar views have been voiced by Rabadán (2005: 156) and more recently by Becher (2010: 9). Tymockzo also implied that corpora can be manipulated, as they are:

> products of human sensibility, connected with human interests and self-interests. All the more reason, therefore, to consider the objects of study, the data gathered in the databases, and the parameters defining the corpora themselves very carefully. (Tymockzo 1998: 3)

It is my proposition that a corpus is only as good as its sampling frame, and the first questions to ask when choosing to use or to compile a corpus is always: “can this corpus answer my research question?” and “what universe of text can findings from this corpus be generalized back to?” Needless to say, there is no magic involved in the compilation and use of corpora, so whatever information is put into the corpus, is the information the researcher will get out. Still, as a

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

reviewer I often come across researchers that seem to hold a very naïve faith in the "objectivity" of corpora, seeming to believe that the mere act of entering texts into the framework of a corpus makes the texts acquire new properties, for instance converting a translated text into an original. I have seen several examples where data from very specific and limited corpora (i.e. data from small corpora of translated or specialized English) have been used to make broad generalizations about the nature of the English language itself. Here we are at the heart of the concept of representativeness, which is one of the core assumptions of corpus linguistics. In fact, representativeness is regarded as a defining factor of a corpus, as can be seen from the following definition from McEnery et al.:

> There are many ways to define a corpus …, but there is an increasing consensus that a corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of particular language or language variety.
> (McEnery, Tono & Xiao 2006: 5) (emphasis original)

The importance of representativeness resides in the fact that data sampled from a representative corpus can be generalized back to a larger universe, which is one of the core uses of quantitative methodology (Johnson 2008). Biber (1993) made this crystal clear:

> The use of computer-based corpora provides a solid empirical foundation for general purpose language tools and descriptions, and enables analyses of a scope not otherwise possible. However, a corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning language as a whole. (Biber 1993: 243)

In order to ensure representativeness, traditionally one would want a variety of text genres as wide as possible when constructing a corpus. However, according to Biber (1993: 243), the most important factor is that the population the texts are drawn from is thoroughly defined and that the entire variability of the population is represented in the corpus.

Consider the following example: The Norwegian Newspaper Corpus, a fascinating monitor corpus created by Knut Hofland[2] that automatically harvests material from 24 Norwegian newspapers daily, is the largest corpus available in Norwegian. It consists of approximately 1.5 billion tokens,[3] and both written varieties (Bokmål

[2] <http://avis.uib.no/avis/om-aviskorpuset/english>

[3] <Avis.uib.no> and personal communication (sms) from Knut Hofland 27.11.17, 10:26.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

and Nynorsk[4]) are represented.[5] Although very useful for the study of many aspects of language change, neologistic usage, lexical productivity and creativity,[6] findings from this corpus cannot be generalized back to the Norwegian language as a whole. Consisting solely of material from newspapers, it does not reflect the entire variability of the Norwegian language in accordance with Biber's definition of representativeness.

This brings us to the next question: does size matter? There is a commonly held misconception that size will even out any problem of representativeness. In a well-designed general corpus, this may well be true, but it does not hold for specialized corpora. As evidenced by the example of the Norwegian Newspaper Corpus, its enormous size cannot even out the fact that it is a specialized corpus that only contains material sampled from one single genre with corresponding genre-specific traits. It is, however, highly representative of the language of Norwegian newspapers.

5. The sampling challenge of the NSPC

How, then, does one go about compiling a corpus that is not only supposed to be representative of the body of text translated from the two Norwegian written language varieties into Spanish, but should also be comparable to an existing English –Spanish Parallel Corpus as well as the CREA reference corpus? The basic conundrum is the following: to be representative, the corpus needs to be sampled from a clearly defined population which results from this corpus can be generalized back to. To be comparable, the sampling frame needs to mirror or be comparable to the corpus it will be compared to, and these two considerations are always competing (Leech 2007: 144; Laviosa 1997.) This daunting task is only briefly described here, as it is described in detail in Hareide & Hofland (2012). In this paper, we sought to make the process of the creation of the NSPC as transparent as possible, by tracing and documenting every single step, the important questions we posed in each of these steps, and the rationale behind the solutions chosen. Our goal was to allow other researchers to answer the all-important first question when choosing a corpus: "Can this corpus answer my research question?" An overview of the NSPC version 1's comparability to the English–Spanish

[4] <http://www.sprakradet.no/Vi-og-vart/Om-oss/English–and-otherlanguages/English/norwegian-bokmal-vs.-nynorsk/>

[5] <http://clarino.uib.no/korpuskel/corpus-list?>

[6] <http://uni.no/en/uni-computing/clu/the-norwegian-newspaper-corpus/>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

subcorpus of the P-ACTRES and the CREA is presented in Table 1. Due to the time frame of the NSPC dating from the year 2000–2009, the Corpes XXI would be the ideal reference corpus, however the NSPC was finalized two years before the Corpes XXI's introduction in October 2013.

Table 1. The comparability of the first versions of the NSPC and P-ACTRES to the CREA corpus, first published in Hareide and Hofland (2012)

Corpora	**NSPC (1.0)**	**P-ACTRES (1.0)**	**CREA**
Number of words	3.01 million	2.5 million	160 million (2008)*
Full texts/excerpts	Full texts	Excerpts	Full texts
Language variety	European Spanish	European Spanish	European Spanish subcorpus
Genres			
Fiction	X	X	X
Non fiction	X	X	X
Popular science	X	X	X
Journalism/editorials	X	X	X
Children's literature	X	X	?**
Magazine articles/press	0	X	X
Theatre	0	0	X
Ephemera	0	X	X

* 50% of the CREA corpus consists of European Spanish.
** The CREA has no children's literature subcorpus.

To ensure maximum comparability with the English–Spanish P-ACTRES corpus (P-ACTRES 1.0), I chose to align my sampling frame with that of the P-ACTRES, with regard to time frame of sampling, the Spanish language variety sampled (European Spanish), the genres represented, size and a focus on non-specialized language "directed at the general public" (Rabadán 2005: 161), as well as size. In order to compile a representative corpus, I first had to define the universe of text the corpus is supposed to be representative of, and to create a strict sampling frame that delimited this population. First, I needed to find out was what kinds of texts are translated from Norwegian into Iberian Spanish. The Norwegian National Library kept a database *called Norbok*[7] where all published translations of texts originally written in Norwegian were registered, and I operationalized the population to "texts featured in the *Norbok* database first published in Norwegian and published in Iberian Spanish between 2000 and 2009" in order to ensure

[7] <http://www.nb.no/baser/norbok2/norbok.php>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

contemporary texts and a timeframe comparable to that of the P-ACTRES 1.0. I made a conscious choice to include both the two Norwegian written standards Nynorsk and Bokmål, and all translators and text types/genres in the population when sampling the texts for the corpus, in accordance with Biber's (1993: 243) principle of maintaining the full variability of the population.

Version 1 of the NSPC described in Hareide & Hofland (2012) consisted of 3.01 million running words, originating from 31 entire text pairs. In 2013, the corpus was expanded to include the entire population of 41 text pairs and now consists of 4.1 million words.[8] Since the corpus is compiled for research in linguistics and translation studies, it is coded for translation relevant meta-data. The comparability between the NSPC and the P-ACTRES is further enhanced by allowing for the subdivision of the NSPC into subcorpora, a strategy that has already been successfully employed by the P-ACTRES 1.0 corpus (Rabadán 2005; Izquierdo, Hofland & Reigem 2008: 36). This means that the Books section of the P-ACTRES 1.0 is comparable to the NSPC's Books section and both corpora are comparable to the Books section of the Spanish CREA reference corpus. The CREA also allows for specified queries based on genre, geography and chronology, where the researcher enters only the desired variables in the query. For an account of the P-ACTRES' comparability to the CREA, see Rabadán (2005: 162). I believe that this procedure ensures the representativeness of both the P-ACTRES and the NSPC while at the same time ensuring their comparability. In addition, the NSPC can also be divided into subcorpora based on parameters that could possibly represent variables in a corpus-based study, such as the author's and the translator's gender, the translator's mother tongue, the text type, and the author's choice of Norwegian written standard (Bokmål or Nynorsk) (Hareide & Hofland 2012: 84–9) to allow for testing of multiple variables. Some researchers, among them De Sutter et al. (2012: 326), criticize research in corpus-based translation studies for not adhering rigorously enough to the corpus linguistic methodology. They claim that researchers in corpus-based translation studies tend to study only one variable, thereby demonstrating insensitivity to other factors such as text type and source-language variation that may influence the variable studied or have explanatory power. Therefore the NSPC can be divided into subcorpora based on parameters that could represent possible variables in a corpus-based study, allowing for testing of

[8] The expansion was funded by the EU under the CIP ICT-PSP programme through the META-NORD project (2011–2013), aimed at creating an open infrastructure to promote the accessibility and reuse of language resources and technologies. Its consortium included organizations from all the Nordic and Baltic countries. Among its main results have been the improvement, linking, documentation, rights clearance, licensing and sharing of many language resources via the META-SHARE catalogue and repository, thereby making the resources more readily available for research and development.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

multiple variables as suggested by de Sutter et al. See Jenset and Hareide (2013) for an example case of a study of multiple variables in the NSPC.

A combination of the NSPC and the P-ACTRES corpora allows for research on two language pairs, and also for comparisons between texts translated from Norwegian into Spanish and from English into Spanish. For an account of the NSPC's comparability to the P-ACTRES and the CREA, see Hareide and Hofland (Hareide & Hofland 2012: 95–97). For an in-depth account of the P-ACTRES' comparability to the CREA, see Rabadán, Labrador & Ramón (2009: 316).

6. Testing of the gravitational pull hypothesis

In Hareide (2017b), I devised a strictly frequency based method for testing Halverson's gravitational pull hypothesis (2003, 2007, 2009, 2010) on two separate language pairs. As one may remember, the test requires independent data sets from two comparable parallel corpora that have the same target language and that are comparable to the same reference corpus. As signaled in Section 4, the test item was the Spanish gerund, and the comparable parallel corpora the P-ACTRES and the NSPC.

To test the hypothesis, I generated two separate data sets, both consisting of a randomized sample of 20% of the sentences containing Spanish gerunds from each text in the P-ACTRES 1.0 (Izquierdo, Hofland & Reigem 2008) and in the NSPC (Hareide & Hofland 2012) as well as their corresponding English and Norwegian source-language sentences. In these two data sets, the Spanish gerunds and the corresponding structures in English and Norwegian that gave rise to the Spanish gerunds in translations were identified. In order to test for prototypicality, the *estar* + gerund structures in the two independent data sets were also identified and analyzed, since this structure has prototypical status (RAE 2009: 2186).

In order to establish empirically whether overlapping resources may constitute direct links in the bilingual's mental lexicon, the English and Norwegian sources of gerunds were examined. No overlapping Norwegian sources were found. The two most frequent English sources are, as predicted, non-finite adverbial clauses and state progressives. This result is concurrent with the predictions of the gravitational pull hypothesis, because the overlap between these resources in English and Spanish seems to constitute a direct link in the bilingual's mental lexicon whereby these structures present themselves as readily available for the translator.

I then went on to pose the following six predictions for empirical testing:

1. The Spanish gerund will be more frequent in Spanish translated from English than in original Spanish. (This hypothesis tests for the overrepresentation of

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

target-language-specific features hypothesis and for all three factors of the gravitational pull hypothesis).

2. In translations from Norwegian, the Spanish gerund will either be underrepresented in accordance with the unique items hypothesis and factor three of the gravitational pull hypothesis, or overrepresented in accordance the overrepresentation of target-language-specific features hypothesis and factor one the gravitational pull hypothesis.
3. The Spanish gerund in particular will be significantly more frequent in translations from English, where overlapping structures exist, than in translationls from Norwegian, where no overlapping structures exist. (This hypothesis tests for evidence of factors 2 and 3 of the gravitational pull hypothesis).
4. The *estar* + gerund structure will be significantly more frequent in Spanish translated from English than in original Spanish. (This hypothesis also tests for factors 1 (prototypicality and frequency in the TL) and 2 (salience or prototypicality in the SL network) of the gravitational pull hypothesis).
5. In translations from Norwegian, the prototypical Spanish *estar* + gerund structure will be overrepresented relative to non-translated Spanish in accordance with factor 1 of the gravitational pull hypothesis and the overrepresentation of target-language-specific features hypothesis – or underrepresented in accordance with factor 3 and the unique items hypothesis.
6. The *estar* + gerund structure will be significantly more frequent in Spanish translated from English than in Spanish translated from Norwegian. This test will only test for evidence of factors 2 and 3 of the gravitational pull hypothesis, but will not test how these three factors interact or their relative strengths.

My study provided support for the first five predictions of the gravitational pull hypothesis on the language pairs Norwegian–Spanish and English–Spanish. This means that the hypothesis correctly predicts the outcome of the translation process regarding frequent and/or prototypical grammatical phenomena, salient structures in the source language, and overrepresentation or underrepresentation as the result of a linkage between the related concepts in the bilingual's mental lexicon. In this study, all these factors pulled towards overrepresentation, including in the two cases where the gravitational pull hypothesis predicted that the outcome could be either overrepresentation or underrepresentation.

Hypothesis six suggests that the estar + gerund structure will be significantly more frequent in Spanish translated from English than in Spanish translated from Norwegian in accordance with factors 2 and 3 of the gravitational pull hypothesis. Surprisingly this hypothesis was falsified, as the prototypical estar + gerund structure is more frequent in translations from Norwegian than from English in my data. This insight could indicate a need to distinguish between frequency

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

and prototypicality in factor one of the hypothesis, as in this case, prototypicality seems to exert greater pull than frequency. One interpretation of the gravitational pull hypothesis is that the more factors pull in the same direction, the stronger the gravitational pull. My study has shown that the interaction between the factors is more complex, and must be studied further, as the initial hypothesis regarding the additive effect of the factors has been refuted in the case of the frequent and prototypical estar + gerund structure (hypothesis 6), but received support in the case of the frequent Spanish gerund (hypothesis 2). This suggests the need for a more complex model that will allow for the study of more variables.

An additional important finding is the fact that the unique items hypothesis was refuted in translations from Norwegian of both the frequent Spanish gerund and the prototypical *estar* + gerund. My study thereby indicates that very frequent or prototypical target-language features appear to be overrepresented in translations in accordance with Baker's (1993) hypothesis, even when they constitute unique items. This finding means that the well attested unique items hypothesis needs to be refined as to when it applies and when it does not, in other words; what is needed for the unique items hypothesis to receive support? One suggestion might be that the unique items hypothesis requires a language pair composed of languages that are very typologically different, such as Finnish (an Uralic language) in contrast to Indo-European languages. Most studies on the unique items hypothesis, such as Tirkkonen-Condit (Tirkkonen-Condit 2002, 2004), Kujamäki (Kujamäki 2004), and Eskola (Eskola 2004), have been conducted on data from the Corpus of Translated Finnish (CTF). However, recent research by Vilinsky (Vilinsky 2012) and Capelle (Capelle 2012) provide support for the hypothesis using the language pairs English–Spanish and French–English respectively, indicating that factors other than typological difference may enter into the equation. For a comprehensive presentation of this project, see Hareide (2017a, 2017b).

Suggestions for the future:

It is my sincere hope that more comparable parallel corpora will be compiled, thereby allowing us to test both simple and complex hypotheses on more than one language pair. I especially hope that my procedure for testing Halverson's gravitational pull hypothesis will be tested on another set of comparable corpora, as replicability is the essence of science. So far the very labour-intensive and extremely costly process of compiling a parallel corpus has represented a very effective hindrance, especially since the various programs needed to compile a parallel corpus involving other languages than English, are not compatible. This means that adaptations to the programs must be done by computer specialists, thereby requiring substantial additional funding if linguists want to compile parallel corpora for instance as part of their PhD project. There is, however, hope that this situation may change in the near future, as Hugo Sanjurjo from the

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

University of León has just presented his PhD on software for compiling parallel corpora, that can be operated by linguists alone. So far a demo version called *The ACTRES Corpus Manager* has been developed, and this framework supports English, Spanish, French and Italian. This prototype allows both compilation and queries of parallel, multilingual and comparable corpora as well as grammatical, semantic and rhetorical tagging in the languages Spanish and English (Sanjurjo Ganzález 2017: 378). I sincerely hope this important work will lead to a surge in comparable parallel corpora, possibly entire families of parallel corpora with comparable sampling frames, thereby allowing us to test our hypotheses on a wide range of language pairs.

7. Conclusions

In this chapter, I have questioned just how scientific the empirical turn in translation studies has turned out to be in practice, and I have addressed the need for the replicability of studies and for well-designed and correctly used corpora. I have introduced and defined the concept of comparable parallel corpora, and highlighted the fact that this innovation allows us to test hypotheses on two language pairs, as well as testing of complex hypotheses like Halverson's gravitational pull hypothesis (Halverson 2010). I have furthermore given a brief overview of the subfield of translation universals of which Halverson's hypothesis forms an integral part, and briefly outlined how the Norwegian Spanish Parallel Corpus (NSPC) (Hareide and Hofland 2012) was complied to be comparable to the English–Spanish P-ACTRES corpus developed at the university of León in Spain (Rabadán 2005, 2007, 2005–2008; Izquierdo, Hofland & Reigem 2008: 31; Rabadán, Labrador & Ramón 2009), before describing how these two comparable parallel corpora were used to successfully test the gravitational pull hypothesis (Hareide 2014, 2017b). I have also drawn attention to Sanjurjo's PhD thesis on the innovative *ACTRES Corpus Manager*, which I sincerely hope will revolutionize the world of Corpus Based Translation Studies by allowing the linguist to construct their own parallel corpora, and hopefully allow us to create entire families of comparable parallel corpora. Just a last word of caution; a sophisticated technical invention like Sanjurjo's software does not solve the problem of badly constructed corpora lacking specific sampling frames, and conclusions drawn wider than the empiric findings can support. If you want to do research in corpus linguistics, you must simply learn the ropes.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Acknowledgment

This chapter draws on and further develops material from three previous works; Hareide and Hofland (2012, Hareide 2014, 2017a and 2017b). There is some overlap between the published versions of these texts. I am grateful for the feedback I have received from my opponents and reviewers in the process of refining the concept of comparable parallel corpora.

References

Altenberg, Bengt & Granger, Sylviane. 2002. Recent trends in cross-linguistic lexical studies. In *Lexis in Contrast. Corpus Based Approaches*, Bengt Altenberg & Sylviane Granger (eds), 3–48. Amsterdam: John Benjamins.

Baker, Mona. 1993. Corpus Linguistics and Translation Studies. In *Text and Technology: in Honour of John Sinclair*, G. Francis & E. Tognini-Bonelli (eds), 233–252. Amsterdam: John Benjamins.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation: Studies in Language Engineering in honour of Juan C. Sagerm* Harold L. Somers (ed.), 175–186. Amsterdam: John Benjamins.

Baker, Mona. 1999. The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4(2): 281–298.

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

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

Blum-Kulka, Shoshana & Levenston, Eddie A. 1983. Universals of lexical simplification. In *Strategies in Interlanguage Communication*, C. Faerch & G Kasper (eds), 119–139. London & New York: Longman.

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Aquisition Studies*, Julianne House & Shoshana Blum-Kulka (eds), 17–35. Tübingen: Gunter Narr.

Capelle, Bert. 2012. English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures* 13(2): 173–195.

Chesterman, Andrew. 2004. Beyond the particular. In *Translation Universals: Do they exist?* Anna Mauranen & Pekka Kujamäki (eds.), 33–49. Amsterdam: John Benjamins.

De Sutter, Gert, Goethals, Patrick, Leuschner, Torsten & Vandepitte, Sonia. 2012a. Towards methodologically more rigorous corpus-based translation studies. In *Across Languages and Cultures* 13(2):137–143.

De Sutter Gert, Goethals, Patrick, Leuschner, Torsten & Vandepitte, Sonia. 2012b. Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures* 13(2): 137–143.

Eskola, Sari. 2004. Untypical frequencies in Translated language. A Corpus based study on a literary corpus of translated and non-translated Finnish In *Translation Universals, Do They Exist?*, Anna Mauranen & Pekka Kujamaki (eds), 83–99. Amsterdam: John Benjamins.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Gass, Susan M. & Larry Selinker. 2008. *Second Language Acquisition: An Introductory Course*. New York. Routeledge.

Goldschneider, Jennifer M. & DeKeyser, Robert. 2005. Explaining the "Natural Order of L2 Morpheme Acquisition" in English: A Meta-analysis of Multiple Determinants. *Language Learning* 55(1): 22–77.

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

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

Halverson, Sandra. 2009. Elements of doctoral training: The logic of the research process, research design and the evaluation of design quality. *The Interpreter and Translator Trainer* 3(1): 79–106.

Halverson, Sandra. 2010. Cognitive translation studies: developments in theory and method. In *Translation and Cognition*, Gregory M Shreve & Erik Angelone (eds). Amsterdam: John Benjamins.

Hareide, Lidun & Hofland, Knut. 2012. Compiling a Norwegian–Spanish Parallel Corpus: methods and challenges. In *Quantitative Methods in Corpus Based Translation Studies*, Michael Oakes & Meng Ji (eds), 75–114. Amsterdam: John Benjamins.

Hareide, Lidun. 2014. Is there Gravitational Pull in translation? A corpus-based test of the Gravitational Pull Hypothesis on the language pairs Norwegian–Spanish and English–Spanish. In *Testing the Gravitational Pull Hypothesis in translation – A Corpus-Based Study of the Gerund in Translated Spanish*. PhD thesis, University of Bergen.

Hareide, Lidun. 2017a. The translation of formal source language lacunas: An empirical study of the overrepresentation of target-language specific features and the unique items hypotheses. In *Corpus Methodologies Explained* Meng Ji, Lidun Hareide, Defeng Li & Michael Oakes (eds) 137–187. London & New York: Routledge.

Hareide, Lidun. 2017b. Is there gravitational pull in translation? A corpus-based study of the *Gravitational Pull Hypothesis* on the language pairs Norwegian–Spanish and English–Spanish. In *Corpus Methodologies Explained* Meng Ji, Lidun Hareide, Defeng Li & Michael Oakes (eds) 137–187. London & New York: Routledge.

Izquierdo Fernández, Marlén. 2008. *Estudio contrastivo y de traducción de las construcciones de -ing inglesas y sus equivalentes en español*, Departamento de Filología Moderna, Universidad de León, León.

Izquierdo Fernández, Marlén. 2012. Corpus-based functionality and translatability: English–Spanish progressive constructions in contrast and translation. *Languages in Contrast* 12 (2): 186–210.

Izquierdo, Marlén, Hofland, Knut & Reigem, Øystein. 2008. The ACTRES parallel corpus: an English–Spanish translation corpus. *Corpora* 3(1): 31–41.

Jenset, Gard Buen & Hareide, Lidun. 2013. A multidimensional approach to aligned sentences in translated text. In *The Many Facets of Corpus Linguistics in Bergen - In Honour of Knut Hofland*, Lidun Hareide, Johansson Christer & Michael Oakes (eds), 195–210. Bergen: Bergen Language and Linguistics Studies (BeLLS).

Ji, Meng, Hareide, Lidun, Li, Defeng Li & Oakes, Michael. 2017. Introduction to Corpus Methodologies Explained An empirical approach to translation studies. In *Corpus Methodologies Explained An empirical approach to translation studies*, Meng Ji, Lidun Hareide, Defeng Li & Michael Oakes (eds), 1–4. London: Routledge.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Malden, Mass.: Blackwell Pub.

Pekka Kujamäki. 2004. What happens to 'unique items' in learners' translation. In *Translation Universals: Do they exist?*, Anna Mauranen y Pekka Kujamäki (eds), 187–204. Amsterdam: John Benjamins.

Laviosa, Sara. 1997. How comparable can 'comparable corpora' be? *Target* 9(2): 289–319.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82*, Stockholm, 4–8 August 1991, Jan Svartevik (ed.), 105–122. Berlin: Mouton de Gruyter.

Leech, Geoffrey. 2007. New resources, ot just better old ones? The Holy Grail of representativeness. In *Corpus Linguistics and the Web*, Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds), 133–149. Amsterdam: Rodopi.

Mauranen, Anna & Kujamäki, Pekka. 2004. Introduction. In *Translation Universals Do they exist?*, Anna Mauranen & Pekka Kujamäki (eds), 65–82. Amsterdam: John Benjamins.

McEnery, Tony, Tono, Yukio & Xiao, Richard. 2006. *Corpus-Based Language Studies an Advanced Resource Book, Routledge applied linguistics*. London & New York: Routledge.

Rabadán, Rosa, Labrador, Belén & Ramón, Noelia. 2009. Corpus-based contrastive analysis and translation universals. A tool for translation quality assessment English --> Spanish. *Babel* 55 (4): 303–328.

Rabadán, Rosa. 2005. Hipótesis, explicaciones y aplicaciones: los caminos de la investigación en traducción inglés-español. In *Estudios de Traducción, Lingüística y filología dedicados a Valentín García Yebra*, 148–170. Arco Libros.

Rabadán, Rosa. 2005–2008. Tools for English–Spanish Cross-Linguistic Applied Research. *Journal of English Studies* 5–6: 309–324.

Rabadán, Rosa. 2007. Divisions, descriptions and applications The interface between DTS, corpus-based research and contrastive analysis. In *Doubts and Directions in Translation Studies*, Yves Gambier, Miriam Shlesinger & Radegundis Stolze (eds), 237–252. Amsterdam: John Benjamins.

RAE. 2009. *Nueva gramática de la lengua Española* / Real Academia Española; Asociación de Academias de la Lengua Española. Madrid.

Sanchez, Tara. 2006. The progressive in the Spoken Papiamentu of Aruba. In *Structure and Variation in Language Contact*, Ana Deumert & Stephanie Durrleman-Tame (eds), 291–314. Amsterdam: John Benjamins.

Sanjurjo González, Hugo. *Creación de un Framework para el Tratamiento de Corpus Lingüísticos*. Phd-thesis, Departamento de Ingeniería Eléctrica y de Sistemas y Automática, *Universidad de León*.

Shlesinger, Miriam. 1991. Interpreter latitude vs. due process. Simultaneous and consecutive interpretation in multilingual trials. In *Empirical Research in Translation and Intercultural Studies* Sonja Tirkkonen-Condit (ed.), 147–155. Tübingen: Gunter Narr.

Snell-Hornby, Mary. 2006. *The turns of Translation Studies. New paradigms or shifting viewpoints?* Amsterdam/Philadelphia: John Benjamins.

Tirkkonen-Condit, Sonja. 2002. Unique Items – over- or under-represented in translated language? *Target* 14(2): 207–220

Tirkkonen-Condit, Sonja. 2004. Unique Items – over- or under-represented in translated language? In *Translation Universals: Do they exist?*, Anna Mauranen & Pekka Kujamäki (eds), 177 – 184. Amsterdam: John Benjamins.

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

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Toury, Gideon. 1991. What are descriptive translation studies into translation likely to yield apart from isolated descriptions. In *Translation Studies: The State of the Art*, Kitty M. Van Leuven-Zwart & Ton Naaijkens (eds), 179–192. Amsterdam: Rodopi.

Tymoczko, Maria. 1998. Computerized Corpora and the Future of translation Studies. *Meta* 43(4): 652–660.

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

Vilinsky, Barbara Martínez. 2012. On the lower frequency of Spanish verbal periphrases in translated texts as evidence for the Unique Items hypothesis. *Across Languages and Cultures* 13(2):197–210.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Living with parallel corpora

The potentials and limitations of their use in translation research

Josep Marco
University Jaume I

Parallel corpora can be used in translation research in at least two ways: as the main source of data or as a supplement to data retrieved from a comparable corpus, enabling data triangulation. In the former scenario, they may throw light on contrastive aspects or on translator techniques and methods. In the latter they will tend to be searched to account for differences perceived between the two components of a comparable corpus. Two case studies will be put forward in order to illustrate these two uses of parallel corpora. Both draw on the English-Catalan subcorpus of COVALT (Valencian Corpus of Translated Literature). The first analyses the translation of meal names whereas the second focuses on the -ment adverb + adjective construction.

Keywords: parallel corpora, main source of data, supplementary source of data, comparable corpora, COVALT

1. Parallel corpora and research on translation: Some landmarks

Let us begin with an obvious fact: parallel corpora were chronologically preceded by monolingual ones. The Brown Corpus, compiled in the 1960s by W. Nelson Francis and Henry Kučera and comprising one million words of American English, was, according to Zanettin (2012: 7), the first electronic corpus in history. In the following decade, under Geoffrey Leech's direction, the Lancaster-Oslo/Bergen Corpus was completed. It was closely modelled upon the Brown Corpus in that it was also made up of a million words – this time of British English. Leech was also the leader of the British National Corpus (BNC) project, completed in 1994. The BNC contained 100 million words and was a huge success.

These corpora were all monolingual and typically used for the study of language use, as well as for the compilation of grammars and dictionaries. Anderman

https://doi.org/10.1075/scl.90.03mar

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

© 2019 John Benjamins Publishing Company

and Rogers (2008: 13) detected the earliest signs of an alliance between electronic corpora and the study of translation in the Scandinavian countries in the 1980s. Paradoxically, in a way, the earliest translation-oriented corpora were not parallel but comparable. Gellerstam (e.g. 1986) compiled a corpus of novels translated into Swedish from English together with a component of novels written originally in Swedish; on this basis, he "was able to show that the distribution of words in translated texts differs from that in original texts, casting new light on the hoary chestnut of 'translationese'" (Anderman & Rogers 2008: 22–23). This line of work was to gain momentum and become mainstream in the 1990s, when Baker published several programmatic articles advocating the use of comparable corpora to uncover the (potentially) universal features of translation. We will come back to this in the next section.

Parallel corpora compiled for the study of contrastive and translation issues did not appear on the scene until the 1990s. The earliest successful attempts are arguably the English–Norwegian Parallel Corpus (ENPC), built in Oslo and Bergen under the joint direction of Stig Johansson and Knut Hofland, and the English–Swedish Parallel Corpus (ESPC), developed at around the same time. Both are parallel bidirectional corpora, as they are made up of pairs of source and target texts in both directions. Therefore, they can function both as parallel and comparable corpora. The ENPC served as a source of inspiration for similar projects elsewhere, such as CEXI or COMPARA. CEXI (Translational English Italian Corpus) "did not result in the creation of a corpus" (Zanettin 2012: 49) due to problems with funding and copyright permissions. COMPARA is a Portuguese –English parallel bidirectional corpus developed from the late 1990s onward which can be freely accessed online (see, for instance, Frankenberg-Garcia & Santos 2003).

These are, in a nutshell, some landmarks in the evolution leading up to the translation-oriented use of parallel corpora. In the following section, we will look at the pros and cons of parallel corpora as translation research tools.

2. Parallel corpora and the study of translation: Potentials and limitations

2.1 Parallel vs. comparable corpora?

Translators are sure to welcome parallel corpora wherever available, as they provide a wealth of *actual* translation solutions at the touch of a button. The advantages of parallel corpora over bilingual dictionaries are obvious: the translation equivalents they offer have actually been used by someone (usually a professional translator) in a specific context, and there are as many of them as matches for a given query. Not for nothing are parallel corpora also known as *translation* corpora in the literature

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

(e.g. Johansson 2007: 5). But how about translation scholars? Do they also regard them as a kind of panacea for the study of translation?

The answer to that question is less straightforward than would appear at first sight. In fact, it largely depends on the aims of the research and its theoretical underpinnings. If a researcher is interested in the relationships between STs and TTs (at whatever level), they may be just as enthusiastic about parallel corpora as we have assumed professional translators to be. But if the researcher is interested in something else, for example what characterizes translated language as opposed to non-translated language, then not necessarily. As hinted in the previous section, that is the kind of research promoted by Baker (e.g. 1993, 1995, 1996) in the 1990s. Her research programme may be summarized in the following, often-quoted lines (1993: 245–246):

> Take a corpus of translations into L from a large number of languages and compare it with a corpus of texts originally written in L, looking for evidence of feature F. Do this for as many Ls as possible. If it is found, for each pair of translation corpus and non-translation corpus, that evidence for F occurs more frequently in the corpus of translated text, then we will have cause to believe that it does so as a result of the translation process and not because of any relationship between any language pair. We may then be justified in calling F a translation universal.

Baker's programme was very influential, and became a source of inspiration for other researchers. But it leaves no room for source texts, and parallel corpora are based on the relationship between target texts and source texts. In fact, the methodological cornerstone of Baker's proposal, as evinced in the lines just quoted, is the monolingual comparable, not the parallel corpus. Exclusion of source texts might be justified as a logical consequence of the programme's theoretical underpinnings: it is target-oriented, along the lines laid by Toury (e.g. 1995). But Toury claimed that, in the descriptive study of translation, target texts come first, not that source texts do not come into the picture *at all*.

There were reactions, criticisms, qualifications. Most of them did not aim to question the validity or, indeed, the relevance of Baker's approach but to supplement it. Kenny, who could judge from a vantage point, in her double capacity as Baker's disciple and user of a parallel corpus in her own research (e.g. 2001), acknowledged a certain resistance to using parallel corpora in translation studies. This was partly due to "the practical difficulty of amassing and aligning source and target texts in electronic form"; but "another, more subtle obstacle could be a reluctance among researchers to fall back into the kind of source text-oriented approach that dominated translation studies until the 1980s" (Kenny 2005: 155). Kenny 2005 quotes Stewart (2000) on the *dethroning* of the source text:

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

> The enthusiasm for the notion that translation should first and foremost respect TL conventions has given the source text (…) at best a subordinate role and at worst something of a raw deal. The SL, once the sovereign of translation theory and now little more than a tiresome interloper and a perennial nuisance, consistently eluding the clutches of tidy theories and neat taxonomies, has – at least for the present – been dethroned.

But Stewart, Kenny claims Kenny 2005: 155, might be "overstating the case", as proponents of the comparable corpus methodology have always warned that one source of explanations for the patterns found (the source text) was missing in their data. She goes on to add that parallel corpora may be used to seek explanations for the findings of studies based on comparable corpora, and that bringing the source text into the picture again by means of parallel corpora, when used in this way, can hardly be regarded as "reactionary" (Kenny 2005: 156). This rather strong language suggests that theoretical positions may carry ideological overtones.

Other voices calling for the (qualified) reinsertion of the source text include Bernardini (2005, 2007). She argues that Baker's method is explicitly intended to address "the elucidation of the nature of translated text as a mediated communicative event"; but "it is the very nature of translation as a mediated communicative event (Baker 1993) that makes an exclusively target-oriented approach to translation analysis methodologically questionable" (Bernardini 2005: 6). This author calls for the balance to be redressed, so that the standard methodology combines comparable with parallel and reference corpora. The data furnished by comparable corpora may be supplemented and sometimes even accounted for (cf. Kenny) by an analysis based on parallel corpora, as many translation decisions derive from the make-up of the source text.

Parallel and comparable corpora, therefore, are by no means mutually exclusive in translation research. Using one type or the other will largely depend on the researcher's aims; but both types may be most fruitfully combined, as their results often enrich and supplement each other. In the following section, attention will be turned to the advantages of parallel corpora as perceived by translation scholars, and the particular uses they have been or may be put to.

2.2 Advantages and uses of parallel corpora

Some scholars, though by no means all, regard parallel corpora as potentially fruitful both for contrastive and translation studies, and envisage some kind of interface between the two disciplines. Malmkjær (1998: 2) addresses the question whether (and the hope that) the use of parallel corpora may narrow the gap between comparative linguistics and translation studies. Even though admitting the risk of overgeneralization, Malmkjær saw, at her time of writing, "disaffection

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

bordering on hostility among some translation scholars with regard to linguistics" and "also a degree of indifference to translation among some linguists" Malmkjær 1998: 2. Linguists, according to some translation scholars, do not understand the nature of translation, a process which, "as translators see it, is a far cry from the introspective method of eliciting isolated equivalents from bilinguals (who are not usually translators) which comparative linguists tend to employ" (Malmkjær 1998: 2). By relying on parallel corpora and, therefore, avoiding introspection, linguists would be in a better position to understand the translation process and collaborate with translation scholars.

Johansson (2007) places more emphasis, instead, on accessibility of meanings. Translation corpora (as he calls them) contain texts which are intended to express the same meanings and have the same discourse functions (2007: 9); paradigms of correspondences can then be established, starting either from the source or the target pole. It is on the basis of these correspondences that meaning patterns can be discerned (Johansson 2007: 28):

> In monolingual corpora, it is relatively straightforward to study forms and formal patterns, but meanings are less accessible. One of the most fascinating aspects of multilingual corpora is that they can make meanings visible through translation.

These patterns, however, are not always clear, and what they often disclose is ambiguity and vagueness (Johansson 2007).

As to the uses that parallel corpora have been put to in translation studies, a brief overview will be presented in the following lines, which will then lead to some kind of generalization.

Malmkjær (1998: 2) claims that the method based on parallel corpora can "promote sense-disambiguation, and can help to identify translation norms and to create machine translation programmes and bilingual dictionaries". Of these three uses, the first is perhaps akin to the interests of contrastive studies, the second to descriptive translation studies and the third to the applied branch of our discipline. There is still another applied use of parallel corpora mentioned by Malmkjær (1998: 3): language learning/teaching and translator training. Most of these uses are subsumed under the two main goals identified by Zanettin (2000: 106): parallel corpora

> can be used by researchers to describe what translators actually do with texts and how they transform them in the process of translation, and they can help practitioners to make informed choices based on translation traditions and norms while translating or learning to translate.

Again, the first goal is descriptive and the second applied, whether parallel corpora are used by professional or trainee translators. Olohan (2004: 24) insists on

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

the descriptive uses when she claims that "choices in translation, as reflected in parallel corpora, may be studied to reveal translation strategies and their effects", even though the term *strategy* did not feature in previous accounts. However, Olohan's "strategies" and Zanettin's "what translators actually do with texts and how they transform them" are very close to *shifts*: all these expressions point to the relationship that obtains between source and target texts at the micro-level, that is as concerns particular segments or problems. Calzada (2005: 168) mentions the analysis of shifts as the first goal of corpus analysis. Even though she refers to electronic corpora in general, not parallel in particular, it seems clear that translation shifts can only be analysed through parallel corpora. That is not true to the same extent of the three other goals mentioned by Calzada (2005): distinguishing between apparent (*prima facie*) synonyms in source and target texts, analysing the style of individual translators and lending support (or otherwise) to alleged translation norms (simplification, explicitation, etc.). The first of these takes us back to Johansson's correspondences across languages, which shed light through translation on how apparently equivalent items are mapped onto each other; the second has been pursued both with comparable (e.g. Baker 2000) and parallel (e.g. Bosseaux 2007) corpora; and the third was first promoted within the comparable corpus paradigm favoured by Baker, as seen above, although translation norms (often called *universals*) are increasingly studied through the combined use of parallel and comparable corpora. Finally, McEnery and Xiao (2008) do not mention the description of translated text among the uses of parallel corpora; they just refer, in rather vague terms, to their role in contrastive studies and the practice of translation. As to the former, they claim (2008: 21) that parallel corpora are useful to know how the same content is expressed in two different languages, but they are a poor basis for contrastive studies because translated texts cannot avoid the *translationese* effect. As to the latter, which falls within the applied branch of translation studies, "parallel corpora can help translators and trainees to achieve improved precision with respect to terminology and phraseology and have been strongly recommended for these reasons" (2008: 26).

By way of summary, then, and leaving aside those uses which clearly pertain to applied translation studies, I would argue that parallel corpora can be used with two main goals in mind in translation research:

1. as the main source of data, with a view to analysing translators' choices, or relationships between source texts and target texts, whether it be at the micro-level of segment pairs (where the main concept is that of translation *technique*, otherwise referred to as *strategy*, *procedure* or *shift*) or at the macro-level of whole texts (where the main concept is that of translation *norms*, whose scope can be broadened if referred to as *laws* or *universals*);

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

2. as a secondary source of data to supplement information provided by a comparable corpus. The rationale of this is, as suggested by Kenny (2005) or Bernardini (2005), that the data yielded by the comparable corpus hold descriptive value, especially if they are geared towards the characterization of translated language, whereas the results of parallel corpus interrogation can be regarded as having explanatory value, insofar as the source text is one of the (admittedly many) factors impinging on translators' decisions.

Moreover, generalizations from the analysis of raw corpus data can be drawn with regard to several variables, "translated language/text" being only of them. Other variables are genre, period, field of specialization, author, work, degree of expertise or individual translator, to name only a few.

In the following sections, two case studies will be put forward in order to illustrate these two uses of parallel corpora. Both studies draw on the English–Catalan subcorpus of COVALT (Valencian Corpus of Translated Literature) – 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. The English–Catalan subcorpus comprises 36 English source texts, amounting to 1,201,757 words, and their corresponding target texts in Catalan (1,343,631 words). The comparable corpus used in the second case study is made up of the translated component in the English–Catalan subcorpus just mentioned and a non-translated component, that is a set of narrative works originally written in Catalan intended to be comparable to the translated component in all relevant respects: place of publication (Valencian Community), date of publication (1990–2000), language (Catalan) and genre (narrative fiction). The non-translated component amounts to 1,899,234 tokens. These corpora have been compiled at the Translation and Communication Department, Universitat Jaume I (Castelló, Spain) and can be accessed for research purposes upon request <http://www.covalt.uji.es>.

3. Case study 1: Analysing the Translation of Meal Names with a Parallel Corpus as the Main Source of Data

Food is arguably a fertile area for the investigation of culture-related translation problems. Outside our discipline, the essentially cultural nature of food and eating has been highlighted, for instance, by Montanari (2004: xi–xii), who claims that food *is* culture through all the stages in the process that brings it to our mouths, that is when it is produced, when it is prepared and when it is consumed. Two particular aspects of the relationship between food and culture, in Montanari's

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

account, are of interest to our concerns here. Firstly, the symbolic value of food differs across time and space (2004: 103) – and both kinds of difference are relevant to translation. And secondly, food is culture both in *substance* and *circumstance* (2004: 129), that is its cultural nature concerns not only the foodstuff itself but also everything surrounding it.

Since food-related, culture-specific items include not only foodstuffs proper but also ways of cooking, adjectives describing tastes or textures, places of eating, names of meals and eating and drinking utensils, amongst other things,[1] we will focus on one of these aspects: meal names. Translation correspondences between meal names in English and Catalan are potentially problematic for two reasons: because each set belongs to a given system and there may well be mismatches across systems; and because terms belonging to a certain system are not always univocal or unambiguous but have a potential for polysemy. In the United Kingdom, for instance, people typically have breakfast first thing in the morning before leaving home for work, and they do not have any other proper meal until noon or early afternoon, when they have lunch. When they arrive home in the evening, they have dinner, for example at six, and they may have some light supper before going to bed. However, on Sundays or other holidays, the main meal of the day is typically eaten at noon or in the early afternoon, and it is often called dinner, not lunch. And then there is tea, which is a light mid-afternoon meal usually consisting of tea accompanied by cake, biscuits or sandwiches – but it must also be borne in mind that *tea* is sometimes used to refer to the main evening meal.

In the Catalan-speaking area (and also generally in Spain), on the other hand, many people have a very light breakfast, or even no breakfast at all, and then at, say, 10 in the morning they may have a *baguette* sandwich or something similar. The first morning meal is usually called *desdejuni* and the second *esmorzar* – even though in many areas *esmorzar* is also used for the former. The afternoon meal is usually eaten later than in most European countries (at about two in the afternoon or even later) and it is called *dinar*. And then dinner is also late (between nine and ten in the evening) and it is called *sopar*. Apart from all that, most children (as well as some adults) usually have *berenar*, a light afternoon meal typically consisting of a sandwich or some kind of cake or snack.

As a result of these mismatches across systems and of variation within a given system, translation correspondences are hard to predict. How are the English meal names above and their Catalan *prima facie* equivalents mapped onto each other in our corpus? In the following paragraphs, an account of the findings will be provided by way of illustration of the kind of use a parallel corpus can be put to in order to answer a specific translation-related question.

[1] For an analysis of food-related culturemes based on COVALT, see Oster & Molés-Cases (2016).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Breakfast occurs 89 times in the English-Catalan subcorpus of COVALT, but two instances must be counted out because of misalignments. Of the remaining 87, it is translated 71 times as "desdejuni/desdejunar" ("breakfast/to have breakfast") and 15 times as "esmorzar". It is clear, then, that the former prevails over the latter as a translation solution, and that fact might be put down either to individual translators' preferences or to their geographical origin. In fact, a corpus of Catalan translations published not in Valencia but in Barcelona might well yield different results, as *esmorzar* is much more frequently used there to refer to the early morning meal than in Valencia. There is just one TT segment matching *breakfast* which does not include "desdejuni/desdejunar" or "esmorzar": "hasn't had his breakfast" → "encara no ha menjat res" ("hasn't eaten anything yet"), where reference to any particular meal is avoided through a more general formulation.

Rather surprisingly, "esmorzar" also features prominently (9 times) as one of the translation solutions for *lunch* (which occurs 48 times in all), even though it is much less frequent than the standard equivalent "dinar" (30 times). The occurrence of *esmorzar* in this context can only be accounted for as a case of interference from Spanish, where the closest formal equivalent of *esmorzar*, that is *almorzar*, is often used to refer to the afternoon meal. Other translation solutions for *lunch* include the more general "menjar" (5 times), used as a noun ("food") or as a verb ("eat"); "el migdia" ("noon", once); and "entrepans" ("sandwiches", once), a more particular choice than the ST "packed lunch", which normally contains more than just sandwiches. Finally, there are two instances where *lunch* is used metaphorically as part of the idiom *(be) out to lunch*, paraphrased by the *Cambridge Idioms Dictionary* as "to be behaving in a very strange or silly way". In one instance, it is difficult to identify the matching TT segment for the idiom because it is part of a long enumeration in which not every item gets translated; in the other, it is translated as "guitllada" ("nuts").

Supper (65 occurrences) yields no surprises. It is consistently translated as "sopar" (60 times), except for 4 cases where it is rendered as the more generic "menjar" ("food/to eat"), and one case where the ST segment is paraphrased, with no meal name mentioned: "and I made a hearty supper" → "que vaig devorar amb ganes" ("which I devoured willingly").

Tea occurs 33 times in the corpus in the sense of "a kind of meal" – as opposed to references to the beverage, in a more restricted sense. In 22 instances it is translated as just "te" ("tea"), even though sometimes the meal-related sense is reinforced by the verb *prendre* ("take/have"). On one occasion "tea things" is rendered as "tassa" ("cup"), because the focus was on the objects on the table, not the meal. But, more interestingly, there are as many as 10 instances in which this quintessentially British meal is replaced in the translation with a target culture meal: "berenar" in 7 instances, but even "sopar" ("dinner", twice) and "dinar" ("lunch", once). What we

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

find here, in a nutshell, is a number of solutions which would hold almost opposite positions on the foreignization/domestication cline: the prevailing technique, literal translation ("tea" → "(prendre el) te"), is foreignizing, whereas the 10 cases where a Catalan meal name is used might be regarded as instances of intercultural adaptation (obviously, a domesticating technique), especially when *tea* is rendered as "berenar" – a meal that has no clear counterpart in the source culture. Translators who choose literal translation may be said to be drawing on their prospective readers' potential knowledge of British eating habits; domesticating translators, on the contrary, have to decide which target culture meal would fit *tea* best in the particular context where it occurs – needless to say, a much more complex operation.

The meal name which proves most problematic in translation, as it gives rise to a wide range of TT solutions, is *dinner*. As can be inferred from the account provided above, *dinner* will typically be translated as *dinar* if it refers to the afternoon meal or as *sopar* if it designates the evening meal. But *dinner* can also be applied to the food being eaten rather than the meal itself, as a social occasion, which complicates matters further. *Dinner* occurs 102 times in the corpus and is translated as: "sopar" (57 times), "dinar" (24 times), "menjar" ("to eat/food", 8 times), "queviures" ("viands", twice), "bona taula" (literally, "good table", once), "ranxo" ("grub", once), "menjada" ("meal", once), "àpat" ("meal", once), "els animalets" ("the little animals", once), "rostits" (as a translation of "roast dinner", once) and it is omitted twice. On two occasions, it is paraphrased so as to avoid repetition; and there is one misalignment, which must be counted out. In other words, "sopar" is the most frequent matching TT segment for *dinner*, followed by "dinar", and most of the remaining solutions are more generic terms either for the meal or for the food eaten at it. These findings, with English meal names as queries and their matching Catalan solutions, are summarized in Table 1.

Table 1. Summary of findings for English meal names as queries and their translation solutions in Catalan in the COVALT corpus

	Breakfast	Lunch	Tea	Dinner	Supper
Desdejuni	71	0	0	0	0
Esmorzar	15	9	0	0	0
Dinar	0	30	1	24	0
Berenar	0	0	7	0	0
Te	0	0	22	0	0
Sopar	0	0	2	57	60
Other	1	9	1	20	5
Misalignments	2	0	0	1	0
Total	89	48	33	102	65

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

All in all, it may be said that such an apparently innocuous word as *dinner* may trigger a large number of different translation solutions, which goes to show that possibly it is not as straightforward as it may seem at first sight. And, more interestingly, the most problematic issue from a cultural perspective as regards meal names is perhaps not specific meals (although *tea* is the one that comes closest to the status of a culture-specific item) but the whole meal systems, which are difficult to map onto each other because each system stems from a different set of daily routines and different assumptions about the role played by food and meals in those routines.

The analysis of the translation of meal names illustrates two kinds of different but closely related studies drawing on parallel corpora. The first kind, already mentioned above, aims to determine what translation techniques (or *shifts*, *procedures*, etc.) prevail in the face of a given translation problem. It is illustrated by *tea*, which can be regarded as a culture-specific item in that, as a meal, it has no counterpart in the target culture. Different techniques along the foreignization/domestication cline have been identified for its translation. The remaining meal names under scrutiny are not culture-specific, but they serve to illustrate the second kind of study, which focuses on *semantic mirror images* (Dyvik 2002). This method was developed to create wordnets and thesaurus entries, but it can also be used to capture the kind of variation observable in the translation of any given ST item. Such variation could not have been predicted on the basis of introspection or the senses collected in bilingual dictionaries. In Dyvik's example (2002: 4), the mirror image (or first *t*-image, as it is also called) of the Norwegian word *tak*, which may mean both "roof" and "grip", in English is the set of all the translations of *tak* found in the Norwegian–English parallel corpus: *roof*, *ceiling*, *cover*, *grip* and *hold*. Each of these words could generate its own first *t*-image by going backwards to the Norwegian component and identifying the set of words triggering it in English.

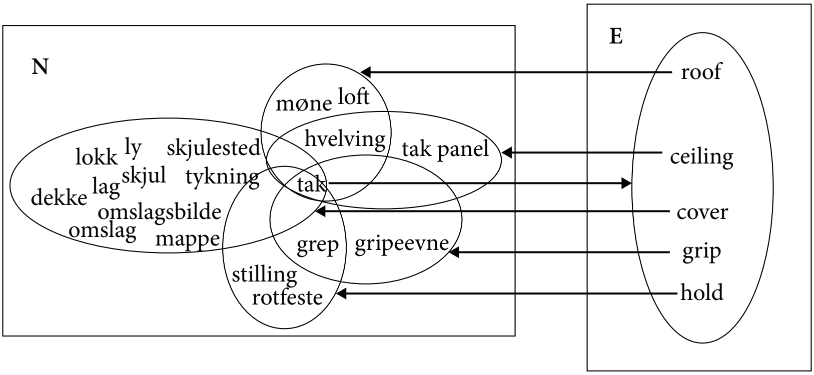


Figure 1. The first and inverse *t*-images of *tak*

This process can be carried out as many times as necessary to create a net of semantic mirrors which would provide a wealth of *bilingual* semantic information. Figure 1 reproduces Dyvik's example.

In our small-scale study, we have derived the mirror images, or *t*-images, of the major English meal names from the English–Catalan parallel corpus. Each item in the set for *breakfast*, for instance (i.e. *desdejuni/desdejunar*, *esmorzar*, etc.) could generate its own *t*-image in English, etc. The kind of semantic information thus derived would be infinitely richer than that provided by bilingual dictionaries.

4. Case study 2: Analysing the construction –*ment* adverb + adjective in Catalan translations with a parallel corpus as a supplementary source of data

Corpus-based translation scholars have paid comparatively little attention to collocation, a key aspect of lexical cohesion which has received fairer treatment in the more general field of corpus linguistics. Some exceptions to this relative lack of interest are Mauranen (2000), Kenny (2001), Baroni & Bernardini (2003), Jantunen (2004), Nilsson (2004), Bernardini (2007) and Dayrell (2007). The research question underlying some of these studies (e.g. Bernardini 2007 and Dayrell 2007) is whether translated texts are more *collocational* than non-translated texts, that is whether the former rely more heavily on fixed lexical associations than the latter. Firth (1957: 179) defined collocation as the company a word keeps. It is beyond the scope of this chapter to review the various problems besetting the concept, especially the "company" part of the definition; but at least the "words" part seems clearer, as it can only be intended to mean word forms or, at most, lemmas. Stefanowitsch and Gries (2003) coined the term *collostruction* in an attempt to expand the notion of collocation and cover "significant associations between words and grammatical structure at all levels of abstractness" (2003: 211). Examples of construction provided by the authors include [X *think nothing of* VPgerund] and the *ditransitive* [S V Oi Od], a predicate with a direct and an indirect object, the latter being remarkably more abstract than the former.

My second case study will focus on a relatively abstract construction in Catalan: manner adverb (ending in the suffix *-ment*) + adjective. Drawing on years of practice in translation, translation analysis and translator training, I would tentatively submit that this construction is less frequent in Catalan than its English counterpart (*-ly* manner adverb + adjective) in English, and that, as a result, it might be overrepresented in Catalan translations from English when compared to Catalan non-translations. If that were the case, it would be an instance of *interference* (Toury 1995) or *shining-through* (Teich 2003), that is the alleged

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

tendency of source texts to leave their imprint on target texts. In order to test this intuition, I will first look at the frequency of occurrence of the construction in both components of my comparable corpus, made up of literary translations into Catalan and Catalan non-translations, all of them published in the 1990s, in order to find out whether there are significant differences. If the construction under scrutiny is significantly more frequent in translations than in non-translations, then I will analyse the source text segments matching the Catalan occurrences of the construction to determine to what extent these occurrences are triggered by its English counterpart. Finally, the English–Catalan subcorpus will be queried from the source pole for the *-ly* manner adverb + adjective construction, with a view to finding out its frequency in the source text component.

The query [(lem = ". + ment") & (pos = "R.*")][pos = "A.*"] was first inserted into the CQP query box to retrieve all cases of lemmas ending in *–ment* which belong to the category "Adverb" followed by an adjective. This was done both for the non-translated and the translated components of the comparable corpus. The query yielded 844 matches for non-translations and 756 for translations. In order to find out whether this difference was statistically significant, the log-likelihood test was applied. The result was highly significant: LL 22.06, p < 0.0001,[2] which means that the statistical significance of the difference can be asserted with a degree of confidence of over 99.99%. The relative frequency of the construction in question, then, is significantly higher in the translated than in the non-translated component of the corpus – it must be borne in mind that the non-translated component (1,899,234 tokens) is considerably larger than the translated one (1,343,631 tokens).

The next step was to manually scan the 756 occurrences of the manner adverb + adjective pattern in the parallel corpus to determine to what extent the higher frequency of this pattern in translations is due to the occurrence of the formally equivalent English pattern in the source text component. Out of these 756 cases, 25 had to be discarded either because they were false matches (i.e. not true representatives of the pattern, for different reasons – for example, a noun ending in *-ment* having been erroneously tagged as an adverb) or because of misalignments. Of the remaining 731 instances, 464 (63.47%) were triggered by the formally equivalent English construction *-ly* adverb + adjective, whereas 267 (36.53%) had other constructions as triggers. That means that almost two thirds of the occurrences of the construction under scrutiny are triggered by the English construction most closely matching it on the plane of formal correspondence. It is a high proportion indeed, but it still leaves room for many other triggers, such as: (a) an adjective

[2] Statistical tests were performed with the University of Lancaster's log-likelihood calculator <http://ucrel.lancs.ac.uk/llwizard.html>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

preceded by a premodifier other than an *–ly* adverb, like *quite* (e.g. "quite unable" → "totalment incapaç" "totally unable"), *a good deal* (e.g. "a good deal puzzled" → "considerablement perplex" "considerably puzzled"), *very*, *downright*, *rather*, *most*, *dammed*, etc.; (b) an adjectival phrase with no premodification in the source text, so that the manner adverb in the Catalan target text may be regarded as an addition, perhaps for emphasis (e.g. "flat" → "completament plana" "completely flat"); (c) two adjectives in succession, rendered as manner adverb + adjective (e.g. "warm-hearted courteous" → "càlidament entranyable" "warmly kind/endearing"); (d) adjective + enough (e.g. "large enough" → "suficientment gran"); (e) a sentence adverbial, that is an adverb acting as adjunct, not as adjective premodifier (e.g. "This certainly is a hard nut to crack" → "Aquesta és realment difícil" "This is really difficult"); and several others. What this means is that, even though no perfect correlation can reasonably be expected in translation data between formally equivalent constructions across languages, the occurrence of our Catalan construction in the parallel corpus seems to be conditioned to a relatively large extent by the occurrence of the *–ly* adverb + adjective construction in the source text.

The third step in the method sketched above was to insert the query "[(lemma = ". + ly") & (pos = "RB.*")][pos = "JJ.*"] in the CQP query box to retrieve all instances of the English construction. The query returned 1,438 matches, which is significantly higher than the number of matches returned for both components of the comparable corpus, either the translated (LL 298.48, $p < 0.0001$) or the non-translated one (LL 546.73, $p < 0.0001$).[3] Tables 2 and 3 summarize the results for frequency of occurrence and statistical differences across corpus components, respectively, as far as the *–ment/ – ly* manner adverb + adjective construction is concerned. Differences are extremely significant in all three possible comparisons (i.e. Catalan translations vs. Catalan non-translations, Catalan translations vs. English source texts and Catalan non-translations vs. English source texts), but the widest difference concerns the latter pair, that is Catalan non-translations as opposed to English source texts. Catalan translations steer a middle course between originals in both languages; they are closer to Catalan originals than they are to English source texts, but even so, as seen above, the difference between Catalan originals and Catalan translations is statistically highly significant.

If we want to interpret these data in terms of the dialectic between the two opposing tendencies of interference and normalization, it can be argued that both tendencies are at play. Starting from the source pole, we are now in a position to state that, out of the 1,438 instances of the *–ly* adverb + adjective pattern in the

[3] In the log-likelihood test, the higher the value, the more significant the difference between the two frequencies. For a 95% level of confidence (i.e. for a p value of < 0.05), the critical value is 3.84; for a 99% level of confidence ($p < 0.01$), the critical value is 6.63; etc.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 2. Frequency of occurrence of the the *–ment/ –ly* manner adverb + adjective construction in the three corpus components

	Raw frequency of occurrence	Relative frequency of occurrence (per million words)
English source texts	1,438	1,196.58
Catalan target texts	756	562.65
Catalan originals	844	444.39

Table 3. Statistical differences across corpus components

	Log-likelihood ratio
Catalan target texts vs Catalan originals	22.06
English source texts vs Catalan target texts	298.48
English source texts vs Catalan originals	546.73

English source texts, only 464 (32.27%) were rendered as manner adverb + adjective in Catalan. In over two thirds of the cases, then, the formally equivalent pattern in the target language was avoided. That might be a sign of normalization: translators are (either consciously or unconsciously) aware that the pattern is more frequent in English than in Catalan and therefore avoid it quite often. However, even if that is the case, source text influence is strong enough for the Catalan pattern to occur more often in translations than in non-translations – a fact that points towards interference, as remarked above.

5. Concluding remarks

It might be concluded from the case studies just presented that parallel corpora show strengths of a different nature depending on whether they are used as the main source of data (case 1) or as a supplementary source of data (case 2). In the former, they enable the researcher to uncover patterns of correspondences between source texts and target texts. Queries may start from the source or the target pole, or they may work both ways. When a single word or a set of semantically related words (e.g. meal names) is explored, semantic mirror images emerge which provide richer bilingual information than dictionaries. When a more abstract translation problem (e.g. culture-specific items) is being investigated, what emerges is a balance of techniques, that is a particular distribution of ST + TT segment pairs across a number of translation techniques. In both cases, light is thrown on what translators actually do when they translate, and the findings of

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

that kind of research are not only interesting in descriptive terms but can also be used in translation practice or exploited in translator training.

As to the second use of parallel corpora (as a supplementary source of data), their main strength lies in their invaluable explanatory potential, as the findings of research exclusively based on comparable corpora are often difficult to explain, no matter how high their value in descriptive terms may be. Using comparable and parallel corpora enables the researcher to triangulate and therefore to provide at least one kind of explanation for the patterns observed in comparable corpora – an explanation derived from source text configuration. In case study 2, for instance, research based on the comparable corpus alone would have revealed that the frequency of use of the –*ment* manner adverb + adjective construction is significantly higher in Catalan translations than in Catalan non-translations; but we could only have *guessed* for an explanation of that fact. On the other hand, research based on the parallel corpus alone would have shown that less than half of the occurrences of the –*ly* adverb + adjective construction in source texts had been translated as its formal correspondence in Catalan; and again we would have been left guessing for reasons for the apparent reluctance of Catalan translators to use this pattern. It is only through the combined use of both types of corpora that the exact balance of the three components comes to light: translators often fail to render the English construction as its closest formal equivalent in Catalan because this construction is much less frequent in original Catalan than in original English (normalization); but even so, they use it much more frequently than original Catalan authors due to source text influence (interference).

There must be many more ways of combining parallel and comparable (and other kinds of) corpora, and it seems reasonable to expect that the corpus-based study of translation will advance through a creative combination of such different resources in the near future.

Acknowledgement

This work was supported by the Spanish Ministry of Economy and Competitiveness [FFI2015–68867–P].

References

Anderman, Gunilla & Rogers, Margaret. 2008. The linguist and the translator. In *Incorporating Corpora. The Linguist and the Translator*, Gunilla Anderman & Margaret Rogers (eds), 5–17. Clevedon: Multilingual Matters.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Baker, Mona. 1993. Corpus linguistics and translation studies – Implications and applications. In *Text and Technology. In Honour of John Sinclair*, Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds), 233–250. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.15bak

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2): 223–243. https://doi.org/10.1075/target.7.2.03bak

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, Harold Somers (ed.), 175–186. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.18.17bak

Baker, Mona. 2000. Towards a methodology for investigating the style of a literary translation. *Target* 12(2): 241–266. https://doi.org/10.1075/target.12.2.04bak

Baroni, Marco & Bernardini, Silvia. 2003. A preliminary analysis of collocational differences in monolingual comparable corpora. <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.3427> (11 April 2017).

Bernardini, Silvia. 2005. Reviving old ideas: Parallel and comparable analysis in translation studies – with an example from translation stylistics. In *New Tendencies in Translation Studies*, Karin Aijmer & Cecilia Alvstad (eds), 5–18. Göteborg: University of Göteborg.

Bernardini, Silvia. 2007. Collocations in translated language. Combining parallel, comparable and reference corpora. In *Proceedings of the Corpus Linguistics Conference*, Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds). <http://ucrel.lancs.ac.uk/publications/CL2007/paper/15_Paper.pdf> (11 April 2017).

Bosseaux, Charlotte. 2007. *How Does it Feel? Point of View in Translation. The Case of Virginia Woolf into French*. Amsterdam: Rodopi.

Calzada Pérez, María. 2005. Corpus electrónicos como herramientas de documentación y formación de traductores. In *La biblioteca de Babel. Documentarse para traducir*, Dora Sales Salvador (ed.), 163–199. Granada: Comares.

Cambridge Idioms Dictionary, 2nd edn. 2006. Cambridge: CUP.

Dayrell, Carmen. 2007. A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics* 12(3): 375–404. https://doi.org/10.1075/ijcl.12.3.04day

Dyvik, Helge. 2002. Translations as semantic mirrors: From parallel corpus to wordnet. In *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)* Göteborg 22–26 May 2002, Karin Aijmer & Bengt Altenberg (ed.), 311–326. Amsterdam: Rodopi.

Firth, John R. 1957. *Papers in Linguistics 1934–1951*. London: OUP.

Frankenberg-Garcia, Ana & Santos, Diana. 2003. Introducing *Compara*, the Portuguese–English parallel corpus. In *Corpora in Translator Education*, Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds), 71–87. Manchester: St. Jerome.

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In *Translation Studies in Scandinavia. Proceedings of the Scandinavian Symposium on Translation Theory (SSOTT)*, Lars Wollin & Hans Lindquist (ed.), 88–95. Lund: Lund Studies in English.

Jantunen, Jarmo Harri. 2004. Untypical patterns in translations: Issues on corpus methodology and synonymity. In *Translation Universals, Do They Exist?* [Benjamins Translation Library 48], Anna Mauranen & Pekka Kujamaki (eds), 101–126. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.48.09jan

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Johansson, Stig. 2007. Seeing through Multilingual Corpora. *On the Use of Corpora in Contrastive Studies* [Studies in Corpus Linguistics 26]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.26

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation. A Corpus-Based Approach*. Manchester: St. Jerome.

Kenny, Dorothy. 2005. Parallel corpora and translation studies: Old questions, new perspectives? Reporting *that* in Gepcolt: A case study. In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, Geoff Barnbrook, Pernilla Danielsson & Michaela Mahlberg (eds), 154–165. London: Continuum.

Malmkjaer, Kirsten. 1998. Love thy neighbour: Will parallel corpora endear linguists to translators? *Meta: Translators' Journal* 43(4): 534–541. https://doi.org/10.7202/003545ar

Mauranen, Anna. 2000. Strange strings in translated language: a study on corpora. In *Intercultural Faultlines. Research Models in Translation Studies, I: Textual and Cognitive Aspects*, Maeve Olohan (ed.), 119–141. Manchester: St. Jerome.

McEnery, Tony & Xiao, Richard. 2008. Parallel and comparable corpora: What is happening? In *Incorporating Corpora: The Linguist and the Translator*, Gunilla Anderman & Margaret Rogers (eds), 18–31. Clevedon: Multilingual Matters.

Montanari, Massimo. 2004. *Il cibo come cultura*. Bari: Laterza.

Nilsson, Per-Ola. 2004. Translation-specific lexicogrammar? Characteristic lexical and collocational patterning in Swedish texts translated from English. In *Translation Universals, Do They Exist?* [Benjamins Translation Library 48], Anna Mauranen & Pekka Kujamaki (eds), 129–141. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.48.11nil

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London: Routledge. https://doi.org/10.4324/9780203640005

Oster, Ulrike & Molés-Cases, Teresa. 2016. Eating and drinking seen through translation: A study of food-related translation difficulties and techniques in a parallel corpus of literary texts. *Across Languages and Cultures* 17(1): 53–75. https://doi.org/10.1556/084.2016.17.1.3

Stefanowitsch, Anatol & Gries, Stefan T. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Stewart, Dominic. 2000. Poor relations and black sheep in translation studies. *Target* 12(2): 205–228. https://doi.org/10.1075/target.12.2.02ste

Teich, Elke. 2003. *Cross-Linguistic Variation in System and Text*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110896541

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.4

Zanettin, Federico. 2000. Parallel corpora in translation studies: issues in corpus design and analysis. In *Intercultural Faultlines. Research Models in Translation Studies, I: Textual and Cognitive Aspects*, Maeve Olohan (ed.), 105–118. Manchester: St. Jerome.

Zanettin, Federico. 2012. *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Working with parallel corpora

Usefulness and usability

Rosa Rabadán
University of Leon

Although parallel corpora are vital for cross-linguistic and natural language processing (NLP) research, most have been designed for just one particular purpose, which may unnecessarily restrict their usefulness and usability. My argument is that the usefulness of existing parallel corpora increases exponentially when data so obtained are combined with those yielded by comparable and/or monolingual corpora. Usability criteria such as the choice of processing tools and adherence to international standards, among others, also have an impact on corpus usefulness. This chapter proposes courses of action that serve to improve the recycling and reprocessing of available resources. It also presents a corpus-based, post-editing and quality assessment application as an illustration of the multifarious uses parallel corpora may serve.

Keywords: parallel corpora uses, parallel corpora reusability, parallel corpora applications

1. Introduction

Parallel corpora are a crucial tool in cross-linguistic studies. Designing a parallel corpus is a costly and time-intensive activity that comprises collecting, annotating, aligning texts, and selecting an efficient corpus manager. Current corpora design decisions are generally guided by the use the corpus is going to have, for example dictionary-making, grammatical contrast, machine translation, machine learning, unveiling translation regularities, etc. More often than not, each new project creates a new parallel corpus from scratch.[1]

[1] One of the reviewers rightly notes that some resources, like Europarl, crop up on different websites, which means that the same collection of documents is being used repeatedly rather than new, diversified repertoires. It is also a fact that, often small-scale, research programs opt to create new parallel corpora, instead of relying on multilingual documents where it is

https://doi.org/10.1075/scl.90.04rab
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

However, a review of a number of parallel corpora featuring Spanish and the way(s) they are used suggests that these custom-made resources have a large, unexploited potential that needs to be identified and assessed.

Poor usability can – and does – prevent corpora from being reused. Efficiency and reasonable accuracy when querying a corpus requires as much standardization as possible. Reliability concerning not only content but also the availability and ease of access to the corpus are key considerations when selecting which corpus to use or how to build a new one.

This chapter provides an argument for collaborative efforts to reuse, reprocess and/or enlarge existing parallel corpora. To this end, Section 2 addresses the broad and narrow definition of parallel corpora and reviews a number of concepts dealing with the usefulness and usability of corpora.

Section 3 presents a number of parallel resources that include at least one of the languages of Spain in its composition. The selection focuses on those corpora that are either openly accessible or from which information is more readily available.

Section 4 reviews the uses of parallel corpora and Section 5 provides a needs analysis of multipurpose bi/multilingual corpora. Section 6 addresses the classic *building or using* dilemma and the challenges that arise from it.

Section 7 presents a corpus-based computerized application for post-editing and translation assessment. Section 8 offers suggestions to parallel corpus users and builders in the form of an action point checklist. The concluding thoughts in Section 9 bring the argument to a close. This chapter relies heavily on the long-term experience of ACTRES as corpus builders and users.[2]

2. Concepts

When applied to language corpora, the term "parallel" can be defined from a broad or a narrow perspective. Broadly, it refers to a bilingual or multilingual collection of original, non-translated computerized texts that serve the same function(s), that is, are used for the same purposes in each language. According to a narrow definition, it refers to a bilingual or multilingual collection of computerized SL and target language (TL) texts, the latter of which are translations of the SL texts

unclear which text is the source text (and language) or whether both texts are the result of multiple translations in both directions– plus, presumably, post editing. Moreover, available parallel collections feature language domains and genres that do not necessarily cater to specific research goals.

[2] For further information about the ACTRES research program and group members see <http://actres.unileon.es/?lang=en>

(see footnote 2). Parallel corpora feature at least two languages. Both languages are amenable to computer processing (i.e. tokenization, annotation, etc.), can be queried using the appropriate filters, and a browser is required for easy access to grammatical, semantic, rhetorical or discourse information, as encoded in the corpus. While the broad definition of parallel corpus includes comparable and SL-TL corpora, the narrow definition is restricted to aligned corpora. Alignment software allows the user to "put together" coupled chunks of text in different languages. This defining feature entails a number of additional constraints for both users and developers.

Users fall mainly into two groups: Natural Language Processing (NLP) professionals on the one hand, and linguists and experts in translation studies, on the other. Both groups require useful and efficient parallel corpora capable of meeting their needs. These needs include sourcing empirical data for contrastive analysis L1-L2-Lx, for creating a testing ground for hypothesis verification, and for providing reliable data for content-rich applications. Parallel corpora can also contribute to the improvement of NLP processes by using human-validated materials in machine learning. Other areas include simplifying various applied tasks –for example, bilingual lexicon induction, grammatical evaluation, etc. –and pushing forward automation for restricted domains and genres. All of these are activities where annotated and validated, reliable parallel corpora are imperative.

The following brief review of (some) parallel corpora that include Spanish and/or other languages of Spain provides a starting point for a discussion of the needs of researchers.

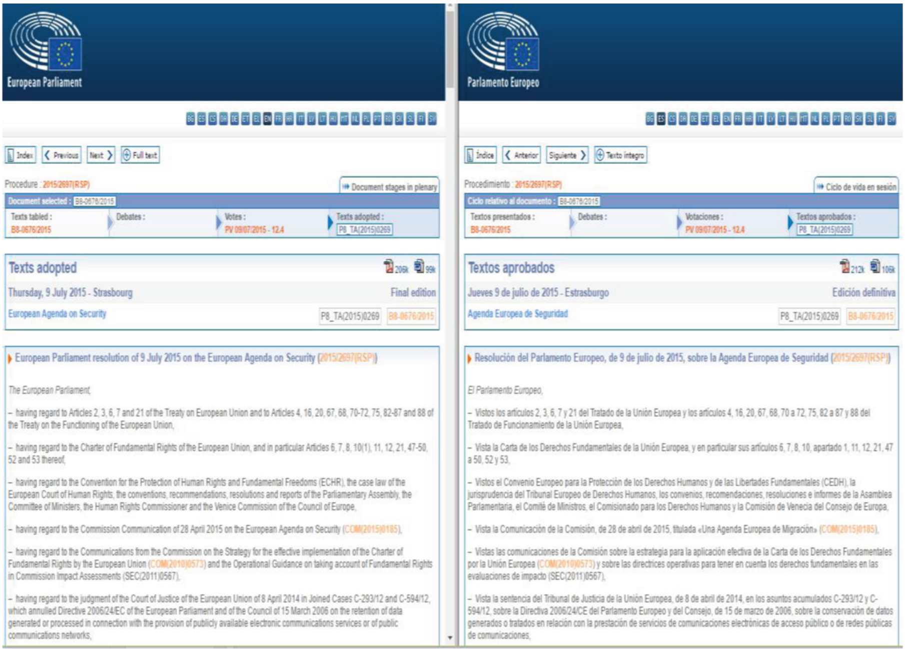
3. Resources

Parallel resources can be organized according to source, namely, government organizations, the industry, and academia

The larger bi/multilingual collections of aligned text are sponsored by multinational–multilingual organizations. The European Parliament (Europarl 2012), the UN (Ziemski, Junczys-Dowmunt & Pouliquen 2016), and officially bi/multilingual countries like Canada (Canadian Hansards, Germann 2017) are an important source of aligned materials sourced from public documentation. These are widely used for a variety of purposes, namely machine translation in the case of the Hansards, or dictionaries (Linguee) and terminological databases (Termium, Iate). However, they present serious drawbacks for researchers in linguistics: it is unclear which of the languages is the SL, the approaches to alignment are widely divergent, and there is a nagging disparity in the way data are measured. Raw quantitative information is presented in terms of sentence pairs, SL or TL

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

tokens, or even paragraphs. Additional issues to be considered when reusing these resources are annotation schemes in the different languages, and the degree of usability, i.e., can they be used and exploited as they are presented? Do they offer usable statistics or, at least, some raw figures about the contents? Is the new user expected to provide additional tools to be able to extract information? Good examples of government corpora are Europarl (Koehn 2005; see Figure 1) or the UN Parallel Corpus (Rafalovitch & Dale 2009).


European Parliament
BG ES CS DA DE ET EL EN FR HR IT LV LT HU MT NL PL PT RO SK SL FI SV
Index
Previous
Next
Full text
Document stages in plenary
Parlamento Europeo
BG ES CS DA DE ET EL EN FR HR IT LV LT HU MT NL PL PT RO SK SL FI SV
Índice
Anterior
Siguiente
Texto íntegro
Ciclo de vida en sesión

Procedure : 2015/2697(RSP)

Document selected : B8-0676/2015

Texts tabled : B8-0676/2015

Debates :

Votes : PV 09/07/2015 - 12.4

Texts adopted : P8_TA(2015)0269

Texts adopted 206k 99k

Thursday, 9 July 2015 - Strasbourg Final edition

European Agenda on Security P8_TA(2015)0269 B8-0676/2015

European Parliament resolution of 9 July 2015 on the European Agenda on Security (2015/2697(RSP))

The European Parliament,

– having regard to Articles 2, 3, 6, 7 and 21 of the Treaty on European Union and to Articles 4, 16, 20, 67, 68, 70-72, 75, 82-87 and 88 of the Treaty on the Functioning of the European Union,

– having regard to the Charter of Fundamental Rights of the European Union, and in particular Articles 6, 7, 8, 10(1), 11, 12, 21, 47-50, 52 and 53 thereof,

– having regard to the Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR), the case law of the European Court of Human Rights, the conventions, recommendations, resolutions and reports of the Parliamentary Assembly, the Committee of Ministers, the Human Rights Commissioner and the Venice Commission of the Council of Europe,

– having regard to the Commission Communication of 28 April 2015 on the European Agenda on Security (COM(2015)0185),

– having regard to the Communications from the Commission on the Strategy for the effective implementation of the Charter of Fundamental Rights by the European Union (COM(2010)0573) and the Operational Guidance on taking account of Fundamental Rights in Commission Impact Assessments (SEC(2011)0567),

– having regard to the judgment of the Court of Justice of the European Union of 8 April 2014 in Joined Cases C-293/12 and C-594/12, which annulled Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks,

Procedimiento : 2015/2697(RSP)

Ciclo relativo al documento : B8-0676/2015

Textos presentados : B8-0676/2015

Debates :

Votaciones : PV 09/07/2015 - 12.4

Textos aprobados : P8_TA(2015)0269

Textos aprobados 212k 106k

Jueves 9 de julio de 2015 - Estrasburgo Edición definitiva

Agenda Europea de Seguridad P8_TA(2015)0269 B8-0676/2015

Resolución del Parlamento Europeo, de 9 de julio de 2015, sobre la Agenda Europea de Seguridad (2015/2697(RSP))

El Parlamento Europeo,

– Vistos los artículos 2, 3, 6, 7 y 21 del Tratado de la Unión Europea y los artículos 4, 16, 20, 67, 68, 70 a 72, 75, 82 a 87 y 88 del Tratado de Funcionamiento de la Unión Europea,

– Vista la Carta de los Derechos Fundamentales de la Unión Europea, y en particular sus artículos 6, 7, 8, 10, apartado 1, 11, 12, 21, 47 a 50, 52 y 53,

– Vistos el Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales (CEDH), la jurisprudencia del Tribunal Europeo de Derechos Humanos, los convenios, recomendaciones, resoluciones e informes de la Asamblea Parlamentaria, el Comité de Ministros, el Comisionado para los Derechos Humanos y la Comisión de Venecia del Consejo de Europa,

– Vista la Comunicación de la Comisión, de 28 de abril de 2015, titulada «Una Agenda Europea de Migración» (COM(2015)0185),

– Vistas las comunicaciones de la Comisión sobre la estrategia para la aplicación efectiva de la Carta de los Derechos Fundamentales por la Unión Europea (COM(2010)0573) y sobre las directrices operativas para tener en cuenta los derechos fundamentales en las evaluaciones de impacto (SEC(2011)0567),

– Vista la sentencia del Tribunal de Justicia de la Unión Europea, de 8 de abril de 2014, en los asuntos acumulados C-293/12 y C-594/12, sobre la Directiva 2006/24/CE del Parlamento Europeo y del Consejo, de 15 de marzo de 2006, sobre la conservación de datos generados o tratados en relación con la prestación de servicios de comunicaciones electrónicas de acceso público o de redes públicas de comunicaciones,

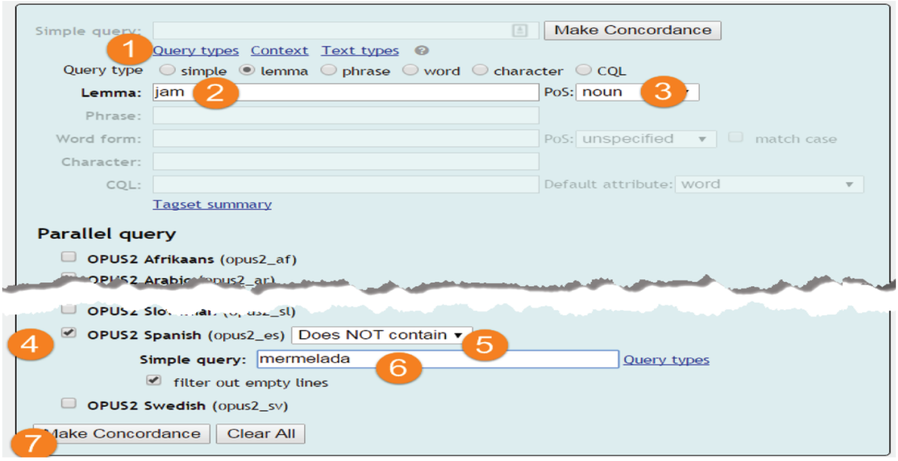
Figure 1. Europarl parallel resources

A second group of parallel resources builds upon translation memories from the industry. Generally, these are a translator's database and they constitute a valuable source of training materials for machine translation and a basis for a number of terminological and lexicographic applications. On average, they are not annotated and cannot be accessed directly because of client-service provider confidentiality. If available, these are good candidates for an annotation upgrade and subsequent reutilization for academic purposes.

Open-source materials mined from the web, including those mentioned above, are the basis of the OPUS project (OPUS 2012). It is a repository of multilingual resources used by SketchEngine to build up parallel/multilingual concordances (Tiedemann 2012). Whenever possible, OPUS corpora offer alignment and POS

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

annotation as shown in Figure 2 (Query: find all sentences in English where *jam* is NOT translated as *mermelada* into Spanish), and 3.[3]


Simple query: Make Concordance

Query types Context Text types

Query type ○ simple ● lemma ○ phrase ○ word ○ character ○ CQL

Lemma: jam PoS: noun

Phrase:

Word form: PoS: unspecified □ match case

Character:

CQL: Default attribute: word

Tagset summary

Parallel query

□ OPUS2 Afrikaans (opus2_af)

□ OPUS2 Arabic (opus2_ar)

□ OPUS2 Slovenian (opus2_sl)

☑ OPUS2 Spanish (opus2_es) Does NOT contain

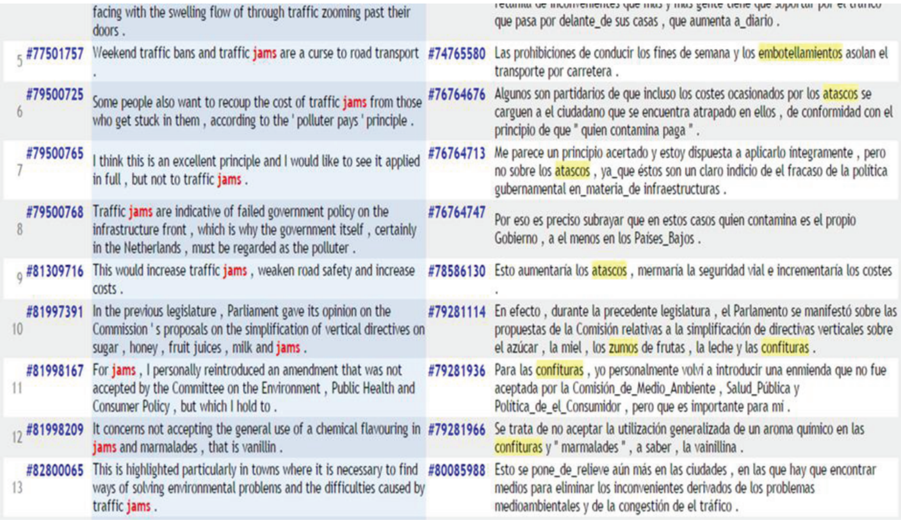
Simple query: mermelada Query types

☑ filter out empty lines

□ OPUS2 Swedish (opus2_sv)

Make Concordance Clear All

Figure 2. Querying OPUS resources using SketchEngine



		facing with the swelling flow of through traffic zooming past their doors .		reclama de inconvenientes que más y más gente tiene que soportar por el tráfico que pasa por delante_de sus casas , que aumenta a_diario .
5	#77501757	Weekend traffic bans and traffic jams are a curse to road transport .	#74765580	Las prohibiciones de conducir los fines de semana y los embotellamientos asolan el transporte por carretera .
6	#79500725	Some people also want to recoup the cost of traffic jams from those who get stuck in them , according to the ' polluter pays ' principle .	#76764676	Algunos son partidarios de que incluso los costes ocasionados por los atascos se carguen a el ciudadano que se encuentra atrapado en ellos , de conformidad con el principio de que " quien contamina paga " .
7	#79500765	I think this is an excellent principle and I would like to see it applied in full , but not to traffic jams .	#76764713	Me parece un principio acertado y estoy dispuesta a aplicarlo íntegramente , pero no sobre los atascos , ya_que éstos son un claro indicio de el fracaso de la política gubernamental en_materia_de infraestructuras .
8	#79500768	Traffic jams are indicative of failed government policy on the infrastructure front , which is why the government itself , certainly in the Netherlands , must be regarded as the polluter .	#76764747	Por eso es preciso subrayar que en estos casos quien contamina es el propio Gobierno , a el menos en los Países_Bajos .
9	#81309716	This would increase traffic jams , weaken road safety and increase costs .	#78586130	Esto aumentaría los atascos , mermaría la seguridad vial e incrementaría los costes .
10	#81997391	In the previous legislature , Parliament gave its opinion on the Commission ' s proposals on the simplification of vertical directives on sugar , honey , fruit juices , milk and jams .	#79281114	En efecto , durante la precedente legislatura , el Parlamento se manifestó sobre las propuestas de la Comisión relativas a la simplificación de directivas verticales sobre el azúcar , la miel , los zumos de frutas , la leche y las confituras .
11	#81998167	For jams , I personally reintroduced an amendment that was not accepted by the Committee on the Environment , Public Health and Consumer Policy , but which I hold to .	#79281936	Para las confituras , yo personalmente volví a introducir una enmienda que no fue aceptada por la Comisión_de_Medio_Ambiente , Salud_Pública y Política_de_el_Consumidor , pero que es importante para mí .
12	#81998209	It concerns not accepting the general use of a chemical flavouring in jams and marmalades , that is vanillin .	#79281966	Se trata de no aceptar la utilización generalizada de un aroma químico en las confituras y " marmalades " , a saber , la vainillina .
13	#82800065	This is highlighted particularly in towns where it is necessary to find ways of solving environmental problems and the difficulties caused by traffic jams .	#80085988	Esto se pone_de_relieve aún más en las ciudades , en las que hay que encontrar medios para eliminar los inconvenientes derivados de los problemas medioambientales y de la congestión de el tráfico .

Figure 3. OPUS querying results

[3] Information about the tagger applied to each language is also available at <http://opus.lingfil.uu.se/trac/wiki/Tagging%20and%20Parsing> (4 July 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

OPUS corpora, as well as parallel materials available through the European Language Resources Association (ELRA 2017),[4] have been designed from an academic standpoint, and include annotation.

Mining parallel texts from the web using, for example, PaCo2 (Parallel Corpora Collector) (San Vicente & Manterola 2012), PT Miner (Chen & Nie 2000), or STRAND (Resnik & Smith 2003) is also a possibility, although these technologies are not yet very efficient.

It is, however, common practice to invest a great deal of effort into designing and building custom-made parallel corpora for particular research projects. These adopt standard practices in compilation, representativeness, and annotation and their use is restricted to project researchers. The Norwegian–Spanish Parallel Corpus (NSPC 2013) (Hareide & Hofland 2012; Hareide 2013) and P-ACTRES 2.0 (Izquierdo, Hofland & Reigem 2008) are illustrative of this practice. Both share technical architecture and were initially designed for human use in linguistic research, specifically, in translation studies and contrastive grammar. Other possible – as yet untested – uses include annotation, language identification, language modelling, bilingual lexicon induction, machine translation, etc.

The ACTRES Parallel Corpus (P-ACTRES 2.0 2018) is a bidirectional English–Spanish corpus consisting of original texts in one language and their translations into the other. P-ACTRES 2.0 contains nearly 6 million words considering both directions together, that is, from original English texts to their Spanish translations (the former P-ACTRES 1.0), and from original Spanish texts to their English translations. (See Sanjurjo-González & Izquierdo in this volume: Section 2.1, Table 1). The materials are distributed into five subcorpora, comprising different text-types: books fiction, books non-fiction, newspaper articles, magazine articles, and miscellaneous texts. Regarding the first two sub-categories, excerpts of around 15,000 words have been extracted from a variety of books. In the case of the other three subcorpora, full articles or texts have been included.

P-ACTRES 2.0 has been annotated with Tree Tagger, and aligned with the *Translation Corpus Aligner* (TCA) (Izquierdo, Hofland & Reigem 2008) (see Figure 4). Querying uses CQP (Evert 2016) and browsing is possible thanks to a browser originally developed for P-ACTRES 1.0 by Knut Hofland (University of Bergen) on the basis of IMS Open Corpus WorkBench (CWB 2013). The browser has since been modified to house both repositories by Hugo Sanjurjo-Gónzalez (University of León) in collaboration with Knut Hofland in order to accommodate

[4] Most ELRA materials must be paid for. There are, however, a number of language resources that can be downloaded free of cost, including the ELRA-W0033 CRATER 2 Corpus and the ELRA-W0023 MLCC Multilingual and Parallel Corpora.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

P-ACTRES 2.0. (See Sanjurjo-González & Izquierdo in this volume: Section 3. Aligning the textual pairs).

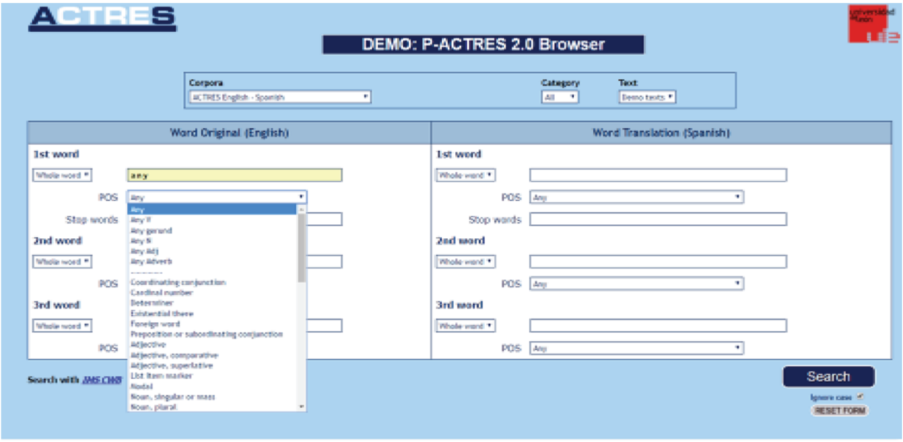


Figure 4. P-ACTRES 2.0 POS categories (TreeTagger)

Many corpus-oriented academic projects have produced their own resource(s). Among those that feature Spanish or any of the languages of Spain are the Multinot corpus (2015), which features POS tagging as well as higher level layers, for example modality for both English and Spanish. BancTrad (2002) is a Multilingual Parallel corpus aligned at sentence level including Catalan, English, French, German, and Spanish. It is POS tagged and includes textual metadata (Badia et al. 2002). COVALT (2005) is a corpus of translated literature. The languages are English, French, German, Spanish, and Catalan. (Marco 2012).

The Laboratorio de Lingüística Informática (2017) has also produced parallel resources, such as the multilingual parallel corpus Arabic-Spanish-English, annotated for named entity recognition (Samy & González-Ledesma 2008) or DIRSI, (DIRectionality in SImultaneous Interpreting), a bilingual English-Italian corpus of transcribed speeches delivered at international conferences and their simultaneous translation performed by professional conference interpreters.

Most of these resources are available to researchers under the modality of *restricted use*. Copyright limitations and the lack of adequate and steady funding for continued technical support, advise corpora builders to be cautious when making their work accessible to the public.[5]

[5] Work to create a hub of accessible resources is underway in the framework of "Red de Excelencia CorpusNet FFI2016-81934-REDT", funded by MINECO (Spanish Ministry of Economic Affairs Competitiveness).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. Uses of parallel corpora

The idea that available parallel corpora need to be exploited more thoroughly instead of embarking upon the costly project of building new ones for each particular research project is also supported by the ample variety of uses for corpora.

Basically, parallel corpora can be used to extract empirical, real-life information about texts and languages used by humans in a wide range of fields, including linguistic modelling (Gilquin 2010), contrastive linguistics (Granger & Petch-Tyson 2003; Rabadán 2010a), translation studies (Ebeling 1998; Halverson 1998), foreign language learning (FLL), and foreign language teaching (FLT) training (Granger & Lefer 2016), and the study of language variation (Biber 1988, 2014), among others.

P-ACTRES 2.0, for example, was initially compiled as a tool to carry out corpus-based grammatical contrastive analysis, but has also been advantageously used, alone or in combination with monolingual corpora, to (a) study translated language, as opposed to non-translated, original language (Rabadán 2011); (b) identify English to→Spanish advanced interlanguage and provide empirical data for targeted remedial work (Labrador 2007); and (c) assess translation quality (Rabadán, Labrador & Ramón 2009).

Data sourced from parallel data can also be used *by computers* to enrich computerized applications aspiring to automatize repetitive (sub)tasks. These include information extraction (IE) (Piskorski & Yangarber 2013), multilingual and cross-linguistic information retrieval (MLIR) (Peters, Braschler & Clough 2012), machine translation (Koehn 2005), machine learning (Pustejovsky & Stubbs 2012), speech recognition (Hu, Isotani & Nakamura 2009), forensic linguistics including determination of origin (Wilson & Foulkes 2014), linguistic fingerprinting, determining authorship, and identifying plagiarism (Coulthard 2004; Prentice, Rayson & Taylor 2012).

The variety of corpus collections and the vastly different approaches from academia, governmental organizations, and industry (TAUS 2016) also suggest that adapting and re-using already existing resources is a more sustainable practice than building a new corpus for each new project. How this might be achieved is discussed in the following sections.

5. Needs analysis

There is a wealth of corpus resources that serve particular projects and interests well. Updated multipurpose corpora, designed to serve a variety of purposes whether in research, education, or the communications industry, tend to be

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

monolingual (e.g. CORPES XXI, COCA) and neglect the particularities of bi/multilingual resources.[6]

As a consequence, bi/multilingual corpora, preferably richly annotated and aligned, are very much in demand. To truly neutralize the "building or using" pitfall (Section 6), these resources need to be reliable, reusable, validated, and standardized, as well as annotated and aligned. In the present situation, achieving this would require collaborative efforts to boost the capabilities of already existing resources, for example, by adding annotation layers (semantic, pragmatic) to currently POS tagged resources, or expanding standardized tagging schemes such as USAS for the semantic layer (Piao, Bianchi, Dayrell, D'Egidio & Rayson 2015), or Multinot for multidimensional annotation (Lavid 2017). Validation and standardization are also critical for the reusability of resources, which imply replicable corpus building protocols and openness when implementing changes leading to improved, high-quality resources.

6. Parallel corpora: Building or using

When deciding whether to use existing parallel resources or instead to embark on the courageous journey of building a new corpus from scratch, it is a good idea to follow a checklist to help identify possible errors or shortcomings in existing corpora. A simple, but effective, initial test regarding the usefulness of available corpora should include at least the following considerations: Does the corpus efficiently address your research problems? Are the contents adequate for your study? Is the compilation reliable? Does it offer sufficient annotation for your purposes? If not, is it possible to upgrade an available resource by, for example, adding an annotation layer, enlarging certain subcorpora, etc.?

The degree of usability required for the user profile should also be assessed. First of all, due consideration should be given to whether the corpus can be queried and browsed as it is or whether it requires additional tools. If it is computationally "self-contained", attention should be paid to the way(s) in which the corpus will be used, the user's linguistic and computing skills, and the sophistication and capabilities of built-in processing and browsing tools.

Whether building or (re)using, the search for maximally processable and reusable parallel resources should take into account design, linguistic, and technical

[6] An additional advantage is that these qualify as *monitor corpora* (Sinclair 2004), that is, corpora that do not have the restriction of a finite limit, either chronological or in terms of size, although components tend to be balanced. The examples provided are also reference corpora, that is, they provide comprehensive information about a given language.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

issues. In terms of design, questions of size (Ghadessy, Roseberry & Henry 2001), balance (Sinclair 2004), and representativeness (Biber 1993) feature high in the common list of concerns of both users and builders of either monolingual or bi/multilingual corpora. Practicalities such as the availability of textual material or directionality acquire greater significance in parallel corpora. The nonexistence of a translation means that there is no "coupled pair" for inclusion in the corpus, which is why taking a target-based approach (Toury 1995) is strongly recommended. On the other hand, the linguistic direction of the materials to be included, for example English into Spanish or Spanish into English, affects both availability and the type of (sub)corpus to be obtained.[7] P-ACTRES 2.0 is bi-directional but, to date, as the availability of suitable materials differ greatly in one and the other direction, it is unbalanced. For example, while there is an abundance of English non-fiction books translated into Spanish, the same is not true for Spanish non-fiction originals translated into English.

Linguistic – and textual – issues that may greatly constrain the selection or the construction of a parallel corpus are the textual mode, genre(s) to be included, domain(s), the language varieties represented in the corpus, and the date of the texts. These also feature high in the concerns of monolingual corpora builders and users, but a parallel corpus perspective brings additional challenges: it may be that there is no cross-linguistic genre correspondence, either because the functions of one particular genre in the SL are distributed among a number of other genres in the TL, or simply because there is no need for such a genre in the target society. An illustrative example is the case of English Directors' Reports (Rabadán, Pizarro & Sanjurjo-Gónzalez 2015), which do not have a counterpart in Spanish as their role is assumed, along with many others, by Spanish companies' annual reports. The representation of language varieties is also a serious issue when dealing with a bi-directional corpus, as diatopic, diastratic, and diaphasic distribution is bound to be different (Coseriu 1981) and also to have a different weight in each of the languages. The selection of the chronological span of the corpus to be used merits even closer attention – in a monolingual corpus this is easily defined, but if we move into parallel corpora the date of the translated texts is as important as the date of their originals. Again, depending on the goal of the project, the translations of one given original, or the works of one particular author, etc., may become the object of study by covering a significant time span in the reception of said texts. However, it may be that empirical information on present-day translation practice(s) or language usage is required, and,

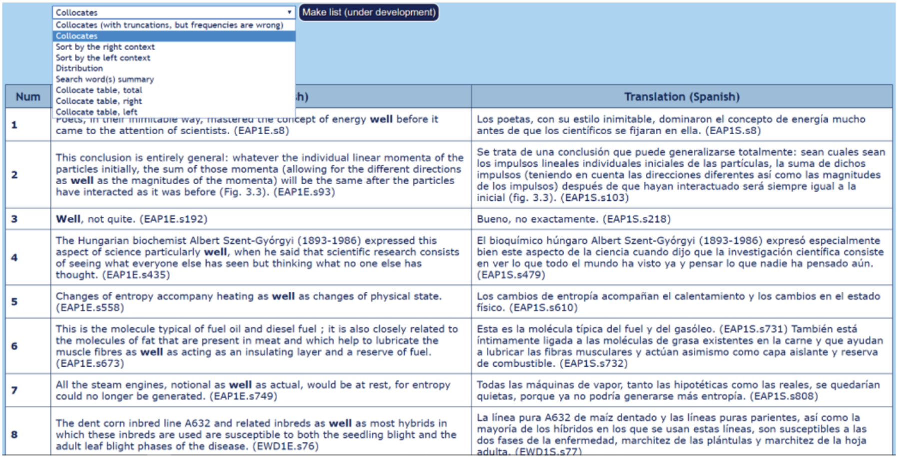
[7] The Index translationum <http://www.unesco.org/xtrans/> and, in the case of Spanish, the ISBN data bases <http://www.mcu.es/webISBN/> are reliable sources for an initial check on source and target availability of materials. Both accessed 8.7.2017.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

if so, a corpus of contemporary originals and their contemporary translations will be necessary.

Technical aspects include alignment, annotation, and usable querying and browsing.

Alignment is a specific concern of parallel corpus builders and solutions come in various standardized formats that may work for a given language pair. The Translation Corpus Aligner (TCA) (Hofland & Johansson 1998) in its different versions is one popular option. It automatically aligns at sentence level, allows interactive correction of alignment (Hofland & Reigem 2017), and underlies the architecture of the English–Norwegian Parallel Corpus (ENPC), its extension the Oslo Multilingual Corpus (OMC), the NSPC, and also of P-ACTRES 2.0 (Figure 5).[8] In the case of the last two, the query processor is CQP and the browser Corpus WorkBench (CQP/ CWB).



Num	h)	Translation (Spanish)
1	Poets, in their inimitable way, mastered the concept of energy **well** before it came to the attention of scientists. (EAP1E.s8)	Los poetas, con su estilo inimitable, dominaron el concepto de energía mucho antes de que los científicos se fijaran en ella. (EAP1S.s8)
2	This conclusion is entirely general: whatever the individual linear momenta of the particles initially, the sum of those momenta (allowing for the different directions as **well** as the magnitudes of the momenta) will be the same after the particles have interacted as it was before (Fig. 3.3). (EAP1E.s93)	Se trata de una conclusión que puede generalizarse totalmente: sean cuales sean los impulsos lineales individuales iniciales de las partículas, la suma de dichos impulsos (teniendo en cuenta las direcciones diferentes así como las magnitudes de los impulsos) después de que hayan interactuado será siempre igual a la inicial (fig. 3.3). (EAP1S.s103)
3	**Well**, not quite. (EAP1E.s192)	Bueno, no exactamente. (EAP1S.s218)
4	The Hungarian biochemist Albert Szent-Gyórgyi (1893-1986) expressed this aspect of science particularly **well**, when he said that scientific research consists of seeing what everyone else has seen but thinking what no one else has thought. (EAP1E.s435)	El bioquímico húngaro Albert Szent-Gyórgyi (1893-1986) expresó especialmente bien este aspecto de la ciencia cuando dijo que la investigación científica consiste en ver lo que todo el mundo ha visto ya y pensar lo que nadie ha pensado aún. (EAP1S.s479)
5	Changes of entropy accompany heating as **well** as changes of physical state. (EAP1E.s558)	Los cambios de entropía acompañan el calentamiento y los cambios en el estado físico. (EAP1S.s610)
6	This is the molecule typical of fuel oil and diesel fuel ; it is also closely related to the molecules of fat that are present in meat and which help to lubricate the muscle fibres as **well** as acting as an insulating layer and a reserve of fuel. (EAP1E.s673)	Esta es la molécula típica del fuel y del gasóleo. (EAP1S.s731) También está íntimamente ligada a las moléculas de grasa existentes en la carne y que ayudan a lubricar las fibras musculares y actúan asimismo como capa aislante y reserva de combustible. (EAP1S.s732)
7	All the steam engines, notional as **well** as actual, would be at rest, for entropy could no longer be generated. (EAP1E.s749)	Todas las máquinas de vapor, tanto las hipotéticas como las reales, se quedarían quietas, porque ya no podría generarse más entropía. (EAP1S.s808)
8	The dent corn inbred line A632 and related inbreds as **well** as most hybrids in which these inbreds are used are susceptible to both the seedling blight and the adult leaf blight phases of the disease. (EWD1E.s76)	La línea pura A632 de maíz dentado y las líneas puras parientes, así como la mayoría de los híbridos en los que se usan estas líneas, son susceptibles a las dos fases de la enfermedad, marchitez de las plántulas y marchitez de la hoja adulta. (EWD1S.s77)

Figure 5. P-ACTRES 2.0 querying results screenshot (CWB browser)

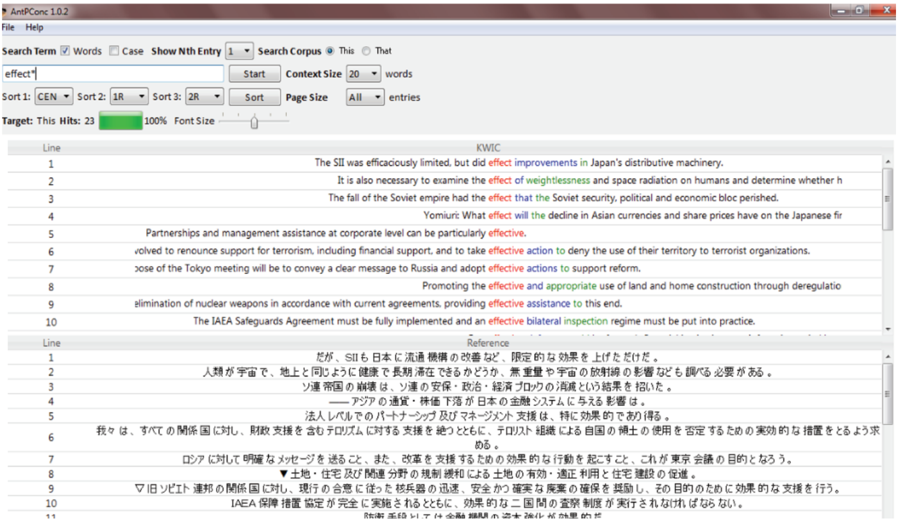
Hunalign, Gargantua or Bleualign are other options, alongside the aligners included in translation memory packages such as SDL Trados Alignment Tool or Wordfast alignment options.

AntPConc (Anthony 2014) (see Figure 6), one of the many freeware tools made available through the author's webpage,[9] is popular among independent corpus builders.

[8] Corpora featuring Norwegian are hosted by Corpuscle (2017), a corpus manager system that is part of the European *Common Language Resources and Technology Infrastructure* network (CLARIN 2012). This integrated system constitutes a source of resources, but it is not directly accessible to independent researchers.

[9] Available at <http://www.laurenceanthony.net/software.html> (19 February 2018).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



Search Term ☑ Words □ Case Show Nth Entry 1 Search Corpus ● This ○ That

effect* | Start | Context Size 20 words

Sort 1: CEN Sort 2: 1R Sort 3: 2R | Sort | Page Size All entries

Target: This Hits: 23 100% Font Size

Line	KWIC
1	The SII was efficaciously limited, but did effect improvements in Japan's distributive machinery.
2	It is also necessary to examine the effect of weightlessness and space radiation on humans and determine whether h
3	The fall of the Soviet empire had the effect that the Soviet security, political and economic bloc perished.
4	Yomiuri: What effect will the decline in Asian currencies and share prices have on the Japanese fir
5	Partnerships and management assistance at corporate level can be particularly effective.
6	volved to renounce support for terrorism, including financial support, and to take effective action to deny the use of their territory to terrorist organizations.
7	pose of the Tokyo meeting will be to convey a clear message to Russia and adopt effective actions to support reform.
8	Promoting the effective and appropriate use of land and home construction through deregulatio
9	elimination of nuclear weapons in accordance with current agreements, providing effective assistance to this end.
10	The IAEA Safeguards Agreement must be fully implemented and an effective bilateral inspection regime must be put into practice.

Line	Reference
1	だが 、SII も 日本 に 流通 機構 の 改善 など 、限定 的 な 効果 を 上げ た だけ だ 。
2	人類 が 宇宙 で 、地上 と 同じ ように 健康 で 長期 滞在 できる か どうか 、無 重量 や 宇宙 の 放射線 の 影響 なども 調べる 必要 が ある 。
3	ソ連 帝国 の 崩壊 は 、ソ連 の 安保・政治・経済 ブロック の 消滅 という 結果 を 招いた 。
4	―― アジア の 通貨・株価 下落 が 日本 の 金融 システム に 与える 影響 は 。
5	法人 レベル での パートナーシップ 及び マネージメント 支援 は 、特に 効果 的 で あり 得る 。
6	我々 は 、すべて の 関係 国 に対し 、財政 支援 を 含む テロリズム に対する 支援 を 絶つ とともに 、テロリスト 組織 による 自国 の 領土 の 使用 を 否定 する ため の 実効 的 な 措置 を とる よう 求める 。
7	ロシア に対して 明確 な メッセージ を 送る こと 、また 、改革 を 支援 する ため の 効果 的 な 行動 を 起こす こと 、これ が 東京 会議 の 目的 と なろ う 。
8	▼ 土地・住宅 及び 関連 分野 の 規制 緩和 による 土地 の 有効・適正 利用 と 住宅 建設 の 促進 。
9	▽ 旧 ソビエト 連邦 の 関係 国 に対し 、現行 の 合意 に 従った 核兵器 の 迅速 、安全 かつ 確実 な 廃棄 の 確保 を 奨励 し 、その 目的 の ため に 効果 的 な 支援 を 行う 。
10	IAEA 保障 措置 協定 が 完全 に 実施 される とともに 、効果 的 な 二 国 間 の 査察 制度 が 実行 さ れ なけれ ば なら ない 。
11	防衛 手段 として は 金融 機関 の 資本 強化 が 効果 的 だ

Figure 6. AntPConc screenshot

Possibly the most troubling among the challenges parallel corpus builders and users face is the need to process text in language-specific ways as the vast majority of processing tools are created and developed for one particular language, generally English.

Parallel corpora necessarily involve at least a second language, and problems arise at every stage of analysis and annotation. If we take POS tagging as an example, the tagsets are not identical, even in the case of so-called language-independent taggers like TreeTagger (Schmid 1994). This makes it necessary to make adjustments when using them as search inputs on aligned source and target texts. The same is true of other types of annotation, namely pragmatic and semantic, as shown by the USAS scheme. Semantic annotation can and must be standardized, but expanding the categories shows that the degree of granularity, meaning range and distribution, are different depending on the language. This is why bi/multilingual semantic annotation uses a simplified, unified taxonomy. Sacrificing degrees of granularity contributes to formulating an increasingly standardized tagset.

A second challenge is the language model underlying the analysis. As underlined in previous works (Rabadán 2016), approaches underlying standard studies of each particular language often represent case studies in contrast in themselves. Different frameworks may yield different (sub)categories, established with slightly different criteria, which result in perfectly equivalent items being considered different phenomena cross-linguistically. One example is English "polarity-sensitive aspectual adjuncts" (Huddleston & Pullum 2002: 710); In the standard grammar of Spanish, these are considered a marginal use of "phasal adverbs" (RAE

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

2009: 2330–2338), and not necessarily related to polarity (Rabadán 2015). The projection of one framework of analysis developed on and for a specific language or family of languages onto another also occurs frequently. The idea that methods and categorization are totally language-independent is not supported by empirical data, as shown by the mapping(s) of modal resources in English and in Spanish. Substantial contrastive evidence suggests that models do not always "travel well" into second languages (Rabadán 2006; Ramón 2009, among others).

A third challenge concerns the querying capabilities of our corpus processing tools. As already mentioned, it is common place for corpora to be designed for a particular purpose. Among the constraints this poses for additional, new research, is the difficulty of querying for different research purposes, which is far from straightforward. An effective way to overcome such difficulties is to "design" a custom-made set of querying strategies that allow already available resources to be used for a new purpose. An example is the querying protocol used to analyze English affixal negation translated into Spanish (Rabadán & Izquierdo 2013).

Nevertheless, available parallel corpora are an unbeatable source of empirical information for content-rich applications, as the following example illustrates.

7. Applications

Increasingly, new application types addressing different applied problems are being added to the more frequent and hugely necessary dictionaries, glossaries or terminological databases. One example of fruitful research based on parallel data is provided by PETRA 1.0© (Rabadán et al. 2014), a post-editing and evaluation aid which uses P-ACTRES 2.0 in combination with the *Corpus de Referencia del Español Actual* (CREA), a monolingual corpus sponsored by the *Real Academia Española de la Lengua*.[10]

PETRA 1.0© is a computerized post-editing tool that uses corpus-based contrastive information to assess the correction and acceptability of grammatical usage in Spanish translations from English. It can also handle poor writing in original Spanish suffering from heavy interference from English. The prototype has been developed for the language pair English→Spanish and the register is *general language* and the mode *written*.[11]

[10] Bowker (2002: 191), for example, proposes combining the data obtained from three different corpora: a quality corpus, a quantity corpus and an inappropriate corpus (a parallel corpus packed with unaudited translations).

[11] The long-term idea is to build up post-editing tools for specific domains and genres, that is, business and management written genres, gadget/ appliance instruction manuals, etc.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

PETRA 1.0© runs on empirical, corpus-based, statistically significant contrastive differences between English and Spanish, which have been conceptualized as *anchors* (Rabadán 2008, 2010), that is, grammatical areas that present different solutions in translated and non-translated Spanish, for example, *cualquier(a)* (Rabadán 2011). As it works on language pair-bound data, PETRA 1.0© would need to activate a different set of *anchors* for each language combination and direction.

The prototype requires the user to upload his or her translation into Spanish and offers two types of evaluation. One is fully automated and uses quantitative information exclusively. The other, more advanced, requires some simple user intervention and uses both quantitative and qualitative data. In both cases, the system uses a built-in, frequency-based, quality control model, which runs on statistical calculations and (i) offers an assessment on a scale between 5 and 0, and (ii) identifies and details the elements that need improvement. The app features a highly usable platform-independent interface that testers have reported as attractive and web-based technology ensures easy access from any device with an internet connection. PETRA 1.0© is primarily addressed to language professionals. Testers have reported high reliability and have given a high rate of approval to the straightforward marking of *areas to improve* and to the possibility of downloading the assessment report as a pdf document.

PETRA 1.0© is constantly evolving and areas of improvement are addressed as they become evident, or as soon as information is reported on malfunctions or possible upgrades. For example, the version 1.1 includes information about the threshold, maximum numbers, and the number of quantitative anchors which could fall within the limits of acceptability.

The 1.0 version is a substantial contribution to the revision toolkit, as well as an empirical tool for dealing with poor TL performance. It saves time and is user friendly, as no particular training or equipment is required, and it offers the possibility of repeated evaluations and the replicability of the evaluation.

While PETRA 1.0© clearly offers advantages to language professionals, from the perspective of the application's builders it presents some drawbacks that could be addressed with collaborative work to identify additional *anchors* to use in assessment. This task demands labor and time-intensive, corpus-based contrast and only a few sub-tasks can be automated, as it necessarily requires expert knowledge.

Other time and money-saving applications based on bi/multilingual parallel corpora include technical and professional writing aids (Labrador, Alaiz-Moretón & Sanjurjo-González 2014) and bitext drafting (Rabadán, Colwell & Sanjurjo-González 2016), among others.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

8. Useful strategies

Rather than revisit and discuss the main argument of this chapter again, that is, the need to increase collaborative efforts to make the most of existing parallel resources, I offer here in summary a short list of "action points" to assist those about to embark on a new project or to update an old one:

1. Have very clear research questions,
2. Review available corpora,
3. Make sure materials are reliable when applicable,
4. Try to build upon what has already been done rather than starting from scratch,
5. If necessary, consider adding extra tags onto already existing corpora,
6. Adopt a creative approach when querying,
7. Use parallel and monolingual corpora in combination to attain richer results,
8. On receiving generous access to resources, respond by contributing generously to those particular resources,
9. Behave and act as part of a community and recall that sharing common interests is the basis of collaborative effort.

9. Conclusions

Parallel corpora are fundamental to a number of tasks and processes of critical importance to linguistics, NLP, and translation studies, among other fields. They provide empirical material for extensive contrastive analysis L1-L2-Lx, which results in deeper insights into the nature and the functioning of both individual languages and cross-linguistic relations.

In spite of the obvious benefits, in the context of the languages discussed in this chapter, this evidence has not sparked a generalized recycling, reusing, and sharing of parallel corpora, for a myriad of reasons. By insisting on (nearly) new resources for upcoming projects, precious time is being wasted and efforts duplicated that could be more profitably invested in upgrading existing corpora to make them amenable to research aims different to those that informed the initial design.

Project outreach and corpus usefulness can also be enhanced by combining parallel data with those yielded by comparable and/or monolingual corpora.

Techniques such as Chesterman's verification of TL fit (2004: 6) can offer valuable insights into interlanguage or discrepancies in obligatory translational adjustments. Reservations about the influence of interlanguage in the results (Mauranen 1999) can be addressed by using bilingual comparable corpora and/or a TL monolingual corpus in the same project.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Adding layers of annotation to already existing corpora is another way of extending their life span. Enriching POS tagged corpora with a semantic annotation layer following a standardized system would provide access to substantial research lines as well as possibilities for practical applications.

For the time being, parallel corpora are a valuable tool for verifying hypotheses, which is fundamental for formulating theoretical principles in fields such as linguistics and translation studies. As sources of validated data, they are crucial for content-rich applications designed to simplify applied tasks for the end user.

Parallel corpora have shown their worth in the improvement of NLP processes using corpus human-validated materials. These are used in machine learning to train larger corpora and expand the range of tasks to be undertaken automatically, among them information retrieval, machine translation, or speech recognition.

And a final caveat: whereas the advent of "big data" has certainly opened fascinating, new possibilities, for the time being the fact remains that only more granular analyses of language can be taken as a reliable basis for pushing forward truly successful language task automation. As yet, our best tool to move beyond the proliferation of project-specific resources is to invest in adaptive corpus practices, and this means verified, richly annotated bi/multilingual parallel corpora which are maximally processable and reusable.

Acknowledgment

Research for his chapter, undertaken as part of ACTRES project [FFI2016-75672-R], has been funded by MINECO, the Spanish Ministry of Economic Affairs and Competitiveness.

References

Anthony, Laurence. 2014. AntPConc (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/> (7 July 2017).

Badia, Toni, Boleda, Gema, Brumme, Jenny, Colominas, Carme, Garmendia, Mireia & Quixal, Martí. 2002. BancTrad: un banco de corpus anotados con interfaz web. *Procesamiento del lenguaje natural* 29: 293–294 < http://www.sepln.org/revistaSEPLN/revista/29/29-Pag293.pdf> (13 November 2018).

BancTrad. 2002. <https://www.upf.edu/es/web/glicom/banctrad> (11 July 2017).

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, Douglas. 1998. *Variation across Speech and Writing*. Cambridge: CUP.

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14(1): 7–34. https://doi.org/10.1075/lic.14.1.02bib

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Bowker, Lynne. 2002. *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Chen, Jian & Nie, Jian-Yun. 2000. Parallel text mining for cross-language IR. In *Proceedings of the 6th International Conference on Computer-assisted Information Retrieval (RIAO 2000)*, 62–77.

Chesterman, Andrew. 2004. Hypotheses about translation universals. In *Claims, Changes and Challenges in Translation Studies*, [Benjamins Translation Library 50], Gyde Hansen, Kirsten Malmkjaer & Daniel Gile (eds), 113. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.50.10pok

CLARIN. European Common Language Resources and Technology Infrastructure. 2012. <https://www.clarin.eu/> (7 July 2017).

COCA. Corpus of Contemporary American English. 201. <https://corpus.byu.edu/coca/> (19 July 2018).

CORPES XXI. Corpus del Español del Siglo XXI. 2016. < <http://web.frl.es/CORPES/view/inicioExterno.view> (19 July 2018).

Corpuscle. 2017. <http://clarino.uib.no/korpuskel/page> (7 July 2017).

Coseriu, Eugenio. 1981. Los conceptos de 'dialecto', 'nivel' y 'estilo de lengua' y el sentido propio de la dialectologia. *Lingüística española actual* 3: 1–32.

Coulthard, Malcolm. 2004. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics* 25(4): 431–447. https://doi.org/10.1093/applin/25.4.431

COVALT. 2005. Corpus Valencià de Literatura Traduïda. <http://cwbcovalt.xtrad.uji.es/cqp-web/> (7 July 2017).

CWB. IMS Open Corpus Workbench. 2013: <http://cwb.sourceforge.net/> (7 July 2017).

Ebeling, Jarle. 1998. Contrastive linguistics, translation, and parallel corpora. *Meta* 43: 602–615. https://doi.org/10.7202/002692ar

ENPC. 1996. English –Norwegian Parallel Corpus. <http://www.hf.uio.no/ilos/english/services/omc/enpc/>(19 July 2018).

Europarl. 2012. Release v7 <http://www.statmt.org/europarl/index.html> (11 July 2017).

European Language Resources Association (ELRA). 2015. <http://www.elra.info/en/catalogues/catalogue-language-resources/> (11 July 2017).

Evert, Stefan. 2016. CQP query language tutorial. CWB Version 3.4 <http://cwb.sourceforge.net/documentation.php> (11 July 2017).

Germann, Ulrich. 2017. Aligned Hansards of the 36th Parliament of Canada release 2001–1a. <https://www.isi.edu/natural-language/download/hansard/> (30 June 2017).

Ghadessy, Mohsen, Roseberry, Robert L. & Henry, Alex (eds). 2001. *Small Corpus Studies and ELT: Theory and Practice* [Studies in Corpus Linguistics 4]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.5

Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions*. [Studies in Corpus Linguistics 39]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.39

Granger, Sylviane & Lefer, Marie-Aude. 2016. From general to learners' bilingual dictionaries: Towards a more effective fulfilment of advanced learners' phraseological needs. *International Journal of Lexicography* 29(3): 279–295. https://doi.org/10.1093/ijl/ecw022

Granger, Sylvianne, Lerot, Jacques & Petch-Tyson, Stephanie (eds). 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi

Halverson, Sandra. 1998. Translation studies and representative corpora: Establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study. *Meta* 43(4): 494–514. https://doi.org/10.7202/003000ar

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Hareide, Lidun & Hofland, Knut. 2012. Compiling a Norwegian–Spanish parallel corpus. Methods and challenges. In *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*, [Studies in Corpus Linguistics 51], Michael P. Oakes & Meng Ji (eds), 75–114. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.51.04har

Hareide, Lidun. 2013. The Norwegian–Spanish parallel corpus, common language resources and technology infrastructure Norway (CLARINO) Bergen Repository <http://hdl.handle.net/11509/73> (4 July 2017).

Hofland, Knut & Johansson, Stig. 1998. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, S. Johansson & S. Oksefjell (eds), 87–100. Amsterdam: Rodopi.

Hofland, Knut & Reigem, Øysten. 2017. Translation Corpus Aligner, version 2. An interactive sentence aligner <http://clu.uni.no/icame/tca2/tca2-abstract.htm> (7 July 2017).

Hu, Xinhui, Isotani, Ryosuke & Nakamura, Satoshi. 2009. Construction of Chinese conversational corpora for spontaneous speech recognition and comparative study on the trilingual parallel corpora. In *ALR7 Proceedings of the 7th Workshop on Asian Language Resources Suntec, Singapore – August 06–07, 2009*. 70–75. Stroudsburg PA: Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1690312&CFID=955832540&CFTOKEN=5706203> (12 July 2017).

Huddleston, Rodney & Pullum, Geoffrey K. 2002. *The Cambridge Grammar of the English Language*.Cambridge: CUP. https://doi.org/10.1017/9781316423530

Izquierdo, Marlén, Hofland, Knut & Reigem, Øysten. 2008. The ACTRES parallel corpus: an English–Spanish translation corpus. *Corpora* 3: 31–41. https://doi.org/10.3366/E1749503208000051

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation, MT Summit, 79–86. <http://www.statmt.org/europarl/> (4 July 2017).

Laboratorio de Lingüística Informática (LLI-UAM). 2017. <http://www.lllf.uam.es/ESP/Recursos.html> (11 July 2017).

Labrador, Belén, Ramón, Noelia, Alaiz-Moretón, Héctor & Sanjurjo-González, Hugo. 2014. Rhetorical structure and persuasive language in the subgenre of online advertisements. *English for Specific Purposes* 34(1): 38–47. https://doi.org/10.1016/j.esp.2013.10.002

Labrador, Belén. 2007. Contrasting ways of expressing restriction in English and Spanish and suggesting translational options into Spanish. *Languages in Contrast* 7(1): 29–52. https://doi.org/10.1075/lic.7.1.03lab

Lavid, Julia. 2017. Annotating complex linguistic features in bilingual corpora: The case of MULTINOT. In *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 201)*, Thierry Declerck & Sandra Kübler (eds), 19–28. Bloomington, IN. <http://ceur-ws.org/Vol-1786/> (11 July 2017).

Marco, Josep. 2012. An analysis of explicitation in the COVALT corpus: The case of the substituting pronoun one(s) and its translation into Catalan. *Across Languages and Cultures* 13(2): 229–246. https://doi.org/10.1556/Acr.13.2012.2.6

Mauranen, Anna. 1999. Will 'translationese' ruin a contrastive study? *Languages in Contrast* 2(2): 161–185. https://doi.org/10.1075/lic.2.2.03mau

Multinot Corpus. 2015. <https://www.ucm.es/funcap/multinot> (7 July 2017).

Norwegian–Spanish Parallel Corpus (NSPC). 2013. <https://repo.clarino.uib.no/xmlui/handle/11509/73> (7 July 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

OMC. Oslo Multilingual Corpus. 2008. <https://www.hf.uio.no/ilos/english/services/omc/> (19 July 2018).

Open Parallel Corpus (OPUS). 2012. <http://opus.lingfil.uu.se/> (7 July 2017).

P-ACTRES 2.0 Corpus. 2018. Demo. <http://actres.unileon.es/?page_id=33&lang=en> (12 November 2018).

Peters, Carol, Braschler, Martin & Clough, Paul. 2012. Cross-language information retrieval. In *Multilingual Information Retrieval. From Research To Practice*, by Carol Peters, Martin Braschler, Paul Clough, 57–84. Berlin: Springer. https://doi.org/10.1007/978-3-642-23008-0_3

Piao, Scott, Bianchi, Francesca, Dayrell, Carmen, D'egidio, Angela & Rayson, Paul. 2015. Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), Denver, Colorado, United States* , 1268–1274. <http://aclweb.org/anthology/N/N15/N15-1137.pdf> (7 July 2017).

Piskorski, Jakub & Yangarber, Roman. 2013. Information extraction: past, present and future. In *Multisource, Multilingual Information Extraction and Summarization*, Thierry Poibeau, Horacio Saggion, Jakub Piskorski & Roman Yangarber (eds), 23–49. Berlin: Springer. https://doi.org/10.1007/978-3-642-28569-1_2

Prentice, Sheryl, Rayson, Paul & Taylor, Paul J. 2012. The language of Islamic extremism: Towards an automated identification of beliefs, motivations and justifications. *International Journal of Corpus Linguistics* 17(2): 259–286. https://doi.org/10.1075/ijcl.17.2.05pre

Pustejovsky, James & Stubbs, Amber. 2012. *Natural Language Annotation for Machine Learning. A Guide to Corpus-Building for Applications*. Sebastopol CA: O'Reilly Media.

Rabadán, Rosa, Labrador, Belén & Ramón, Noelia. 2009. Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment English –Spanish? *Babel* 55(4): 303–328. https://doi.org/10.1075/babel.55.4.01rab

Rabadán, Rosa & Izquierdo, Marlén. 2013. A corpus-based analysis of English affixal negation translated into Spanish. In *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson* [Studies in Corpus Linguistics 54], Karin Aijmer & Bengt Altenberg (eds), 57–82. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.54.05rab

Rabadán, Rosa, Alaiz-Moretón, Héctor, Fernández, Ramón-Ángel, García-Gallego, Ana, Gutiérrez-Lanza, Camino, Labrador, Belén, Ramón, Noelía & Sanjurjo-González, Hugo. 2014. *Procedimiento de evaluación de la calidad gramatical de las traducciones al español de textos en lengua inglesa (PETRA 1.0)* <http://actres.unileon.es/?page_id=50&lang=en>

Rabadán, Rosa, Pizarro, Isabel & Sanjurjo-González, Hugo. 2015. GEDIRE©: A directors' reports writing tool. Paper presented at *CILC 2015. 7th International Conference on Corpus Linguistics*. Valladolid, 5–7 March 2015.

Rabadán, Rosa, Colwell, Veronica & Sanjurjo-González, Hugo. 2016. Bi-texting your food: Helping the gastro industry reach the global market. In *CILC 2016. 8th International Conference on Corpus Linguistics* [EPiC Series in Language and Linguistics 1]. Antonio Moreno Ortiz & Chantal Pérez-Hernández (eds), 361–371.

Rabadán, Rosa. 2006. Modality and modal verbs in contrast: Mapping out a translation(ally) relevant approach English–Spanish. *Languages in Contrast* 6(2): 261–306. https://doi.org/10.1075/lic.6.2.04rab

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Rabadán, Rosa. 2008. Refining the idea of 'applied extensions'. In *Beyond Descriptive Translation Studies: Investigations in homage to Gideon Toury* [Benjamins Translation Library 75], Anthony Pym, Miriam Shlesinger & Daniel Simeoni (eds), 103–117. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.75.09rab

Rabadán, Rosa. 2010. *Applied Translation Studies. In Handbook of Translation Studies 1*, Yves Gambier and Luc van Doorslaer (eds). <https://beta.benjamins.com/online/hts/articles/app1> (7 July 2017).

Rabadán, Rosa. 2010a. English–Spanish contrastive analysis for translation applications. *Quaderns de Filologia. Anejo n.*° 73: 161–180.

Rabadán, Rosa. 2011. Any into Spanish: A corpus-based analysis of a translation problem. *Linguistica Pragensia* 21(2): 57–69. https://doi.org/10.2478/v10017-011-0005-y

Rabadán, Rosa. 2015. A corpus-based study of aspect: Still and already + verb phrase constructions into Spanish. In *Cross-linguistic Studies at the Interface between Lexis and Grammar*, Karin Aijmer & Hilde Hasselgård (eds). *Nordic Journal of English Studies* 14(1): 34–61.

Rabadán, Rosa. 2016. Proposals in meeting minutes: An English–Spanish corpus-based study. *Languages in Contrast* 16(2): 213–238. https://doi.org/10.1075/lic.16.2.03rab

Rafalovitch, Alexandre & Dale, Robert. 2009. United Nations general assembly resolutions: A six-language parallel corpus. In *MT Summit* XII, 292–299. Ottawa: AMTA. <http://uncorpora.org/Rafalovitch_Dale_MT_Summit_2009.pdf> (7 July 2017).

Ramón, Noelia. 2009. Translating epistemic adverbs from English into Spanish: Evidence from a parallel corpus *Meta* 54(1): 73–96. https://doi.org/10.7202/029794ar

Real Academia Española (RAE). 2009. *Nueva gramática de la lengua española*. Madrid: Espasa.

Resnik, Philip & Smith, Noah A. 2003. The web as a parallel corpus. *Computational Linguistics* 29(3): 349–380. https://doi.org/10.1162/089120103322711578

Samy, Doaa & González-Ledesma, Ana. 2008. Pragmatic annotation of discourse markers in a multilingual parallel corpus (Arabic–Spanish–English). *Proceedings of the VI Language Resources and Evaluation Conference (LREC)*. Marrakech, Morocco. <http://www.lrec-conf.org/proceedings/lrec2008/pdf/828_paper.pdf> (7 July 2017).

San Vicente, Iñaki & Manterola, Iker. 2012. PaCo2: A fully automated tool for gathering parallel corpora from the Web. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. <https://www.researchgate.net/publication/230799614_PaCo2_A_Fully_Automated_tool_for_gathering_Parallel_Corpora_from_the_Web> (19 July 2018).

Schmid, Helmut. 1994. TreeTagger – a part-of-speech tagger for many languages. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/> (7 July 2017).

Sinclair, John. 2004. Corpus and text. Basic principles. In *Developing Linguistic Corpora: a Guide to Good Practice. Corpus and Text – Basic Principles*, Martin Wynne (ed.). <https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm> (11 July 2017).

TAUS. 2016. <https://www.taus.net/knowledgebase/index.php/Parallel_corpus> (4 July 2017).

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)* <http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf> (4 July 2017).

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond* [Benjamins Translation Library 4]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.4

Wilson, Paul & Foulkes, Kim. 2014. Borders, variation, and identity: Language analysis for the determination of origin (LADO). In *Language, Borders and Identiy*, Dominic Watt & Carmen Llamas (eds), 218–229. Edinburgh: EUP.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Ziemski, Michał, Junczys-Dowmunt, Marcin & Pouliquen, Bruno. 2016. The United Nations Parallel Corpus v1.0. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 3530–3534. ELRA.<http://www.lreconf.org/proceedings/lrec2016/pdf/1195_Paper.pdf> (13 November 2018).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Innovations in parallel corpus alignment and retrieval

Martin Volk
University of Zurich

In this chapter, we give an overview of parallel corpus annotation, alignment and retrieval. We present standard annotation methods such as Part-of-Speech tagging, lemmatization and dependency parsing, but we also introduce language-specific methods, for example for dealing with split verbs or truncated compounds in German. Our corpus annotation includes the identification of code-switching within sentences as a special case of language identification. We argue for careful sentence and word alignment for parallel corpora. And we explain how word alignment is the basis for a wide range of applications from translation variant ranking to lemma disambiguation.

Keywords: multiparallel corpora, corpus annotation, word alignment, corpus retrieval

1. Introduction

In recent years, an increasing number of large parallel corpora have become available for research in natural language processing. The best known is Europarl, which contains the proceedings of the European parliament (Koehn 2005; Graën et al. 2014) with around 50 million tokens in the languages of the European Union. Other well-known multilingual and multiparallel corpora are JRC Acquis (Steinberger et al. 2006) with the EU law collection, OpenSubtitles (Lison & Tiedemann 2016), United Nations documents (Ziemski et al. 2016), and collections of patent applications (Junczys-Dowmunt et al. 2016).

Switzerland is a country with four official languages (French, German, Italian, and Rumansh) and, because of the many international companies and organizations in Switzerland, English is becoming ever more popular. Therefore, there is a constant need for translations between these languages and this is a natural basis

https://doi.org/10.1075/scl.90.05vol
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

for a plethora of parallel corpora. We have taken advantage of this situation and collected and annotated various Swiss parallel corpora.

Our research group specializes in building parallel corpora for special domains which span over time: We have digitized parallel texts from the Swiss Alpine Club in French and German from 1957 until today (Göhring & Volk 2011), banking texts in English, French, German and Italian from 1895 up to the present (Volk et al. 2016a), and the announcements of the Swiss federal government (DE: Bundesblatt, FR: Feuille fédérale, IT: Foglio federale) since 1849.

We have surveyed different online search systems over parallel corpora, the best known being Linguee, Glosbe and Tradooit (Volk et al. 2014). For each, we identified different user groups (language learners, translators, linguists, language technologists) and their different needs when querying these corpora.

In the current chapter, we will focus on the latest methods in the automatic annotation and alignment of parallel corpora. We will argue that word alignment across languages improves annotation. In fact, we argue that the annotation of parallel corpora will be superior in quality to any monolingual annotation. We focus on parallel corpora for linguistic and translation studies, but we believe that parallel corpus search systems are also interesting for language learners for viewing translation variants in context and for computational linguists who want to evaluate the quality of, for instance, automatic sentence alignment or word alignment.

The chapter is structured as follows. In Section 2, we describe corpus annotation methods, starting from standard Part-of-Speech tagging, lemmatization and dependency parsing. This will lead us to advanced techniques such as prefix re-attachment of split verbs in German and the detection of code-switching within sentences. Section 3 is devoted to alignment techniques and their benefits for corpus annotation such as word sense disambiguation with practical applications in lemma disambiguation and named entity recognition. In Section 4, we give usage examples of our parallel corpora including translation discovery and translation error detection. We conclude with speculations about the future of parallel corpus retrieval systems.

2. Corpus annotations

Corpus building starts with corpus collection, cleaning, and tokenization. The latter is often language-specific and therefore multilingual corpora require language identification. Typically, identification is done on the sentence level. For each sentence, we compute the language in order to be able to use the appropriate processing tools during corpus annotation. Using a part-of-speech tagger that was trained for one language to annotate a sentence in another language will lead to erratic

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

results. Therefore, language identification on the sentence level is of paramount importance for all texts with mixed languages, if only for the proper treatment of quotes and titles in "foreign" languages.

Automatic language identification is typically based on statistics of character-trigrams and is reliable for sentences of length 50 and more characters. For shorter sentences, we usually propagate the language of the previous sentence. This leaves the problem of code-switching within a sentence. Our corpora have "foreign" language fragments within sentences. Example (1) shows a German sentence from our Alpine corpus with an insertion in English framed by quotation marks.

(1) DE: … und ich finde es **"very nice and delightful"** einen Vortrag halten zu dürfen.
EN: … and I find it *"very nice and delightful"* to be able to give a talk.

We have therefore included a code-switch detector in our corpus annotation pipeline which checks the language for intra-sentential word sequences framed by quotation marks. If the language within the frame is different from the language of the surrounding sentence, then we annotate the word sequence as a code-switching block and exempt it from the processing tools (Volk & Clematide 2014). We decided not to assign language-specific POS tags to the code-switch sequence since it often is not a complete sentence. Rather, we assign the POS tags for foreign words in the respective language of the matrix sentence (e.g. FW for English, ET for French, FM for German).

2.1 General corpus annotation

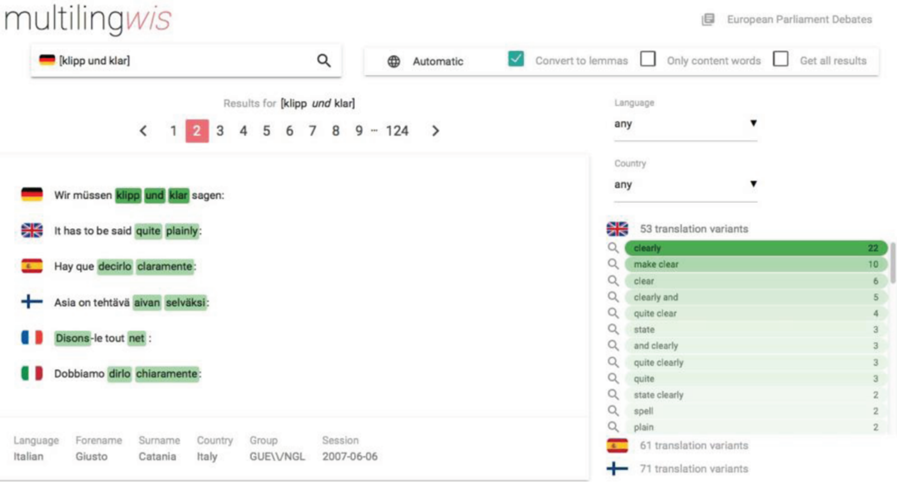
Part-of-speech (PoS) tagging is standard procedure when the corpora are intended for linguistic research. There are a number of POS taggers available with parameter files for many large languages of Europe. Most of the time, the parameter files are the result of training the taggers on newspaper texts. This means that the taggers work best on newspaper texts and gradually worse the more the corpus material differs from newspapers.

In addition, the standard parameter files cater to the "usual" POS tag set for each language. This can be bothersome for the user of a multi-parallel corpus who will have to adapt his queries to each tag set. For example, the tag for proper name in the German STTS is *NE*, in French it is typically *N_P*, and in the English Penn tagset it is *NNP* for singular and *NNPS* for plural names. Therefore, in multilingual environments we welcome universal POS tags (Petrov et al. 2012), which facilitate retrieval enormously. There are now mappings from many language-specific POS tagsets to universal POS tags.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Often, POS tagging also provides lemmas. For example, the TreeTagger outputs lemmas for word form – lemma pairs that it has seen in its training corpus. For all other word forms, the corpus builder may provide a tagger lexicon with additional pairs. These pairs lack the probabilities that tagger training derives from a manually annotated corpus, but for word forms with little or no PoS ambiguity the extension of the tagger lexicon is still useful. So, additional lemma information from other corpora or from dictionaries or morphological analyzers is valuable.

Recent parallel corpora have been annotated with more annotation layers: Named entity recognition (NER) is a popular method for a first step towards semantic annotation. Typically, it involves the recognition of person names, location names and organization names which are the central classes when processing newspaper texts. Special text types may require other name classes (e.g. event names) or more fine-grained distinctions. For example, in our parallel corpus of Alpine texts we sub-classify toponyms into the name classes of mountains, glaciers, lakes, valleys, cabins, and cities. Toponyms are essential for mountaineering reports and therefore appear very frequently.



[klipp und klar]

Automatic ☑ Convert to lemmas □ Only content words □ Get all results

Results for [klipp *und* klar]

Language: any ▾

Country: any ▾

Wir müssen klipp und klar sagen:

It has to be said quite plainly:

Hay que decirlo claramente:

Asia on tehtävä aivan selväksi:

Disons-le tout net :

Dobbiamo dirlo chiaramente:

Language	Forename	Surname	Country	Group	Session
Italian	Giusto	Catania	Italy	GUE\\/NGL	2007-06-06

53 translation variants

clearly	22
make clear	10
clear	6
clearly and	5
quite clear	4
state	3
and clearly	3
quite clearly	3
quite	3
state clearly	2
spell	2
plain	2

61 translation variants

71 translation variants

Figure 1. A Multilingwis query with hits in six Europarl languages

Shallow NER includes only the recognition and classification of names. A deeper analysis includes co-reference resolution (Ebling et al. 2011) and entity linking (sometimes called grounding). Monolingual co-reference resolution will deal with mention variants like *Grand Combin* = *Combin*, while multilingual co-references will catch translation variants as for example *DE: Matterhorn* = *FR: Combin* = *IT: Combino*. Monolingual co-references might include anaphora resolution and will thus allow for investigating coherence phenomena in texts.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Entity linking might seem irrelevant for linguistic research. Why is it important to know whether "Michael Jackson" in a given document relates to the singer or the basketball player? Why should we care whether "Sydney" refers to a town in Australia or in Canada? Obviously, the semantic interpretation of the surrounding text will profit from the disambiguation of these names and is therefore also of interest to linguists and translation scholars. Lately, dependency parsing has become available for many languages (e.g. Maltparser, Spacy, Stanford, etc.). These parsers allow for the efficient analysis of large corpora with a labeled attachment score of 80–90% (McDonald & Nivre 2011) and higher values for unlabeled attachment. Even though parsing is far from perfect, the automatically assigned syntax information opens a whole range of new possibilities for corpus linguistics. For example, searches for verb-object relations no longer need to speculate on co-occurrence in some arbitrary range but can be conditioned on parsing evidence. With parsing, we can find candidates for support verb construction like *to take into consideration* or for verb sub-categorizations for particular prepositional objects like *to wait for*.

In analogy to universal POS tags, there are also universal dependency labels. We encourage the use of such universal labels especially in the context of parallel corpus annotation. The use of language-specific dependency labels only makes sense when fine-grained distinctions that are lost in the conversion to universal labels are being investigated.

Sentiment annotation might be useful for some text types (such as newspaper editorials) to mark whether a paragraph is positive or negative about an event, an organization or a product. Tools for sentiment analysis have been developed for social media texts (tweets, blogs, reviews), but can also be used to label other text types. Annotations of parallel texts in multiple languages profit from each other through cross-language comparisons. If the text is positive in language 1, its translation will also be positive in language 2. So, sentiment analysis tools for both languages can profit from each other, and the precision of annotations can be increased by using only judgments that are in agreement.

2.2 Exploiting parallel corpora for annotation

Traditionally, most corpus annotation is done monolingually. This means that POS tagging, lemmatization and parsing of for example a German corpus is done irrespective of a parallel text in English or any other language. However, the parallel text may help to disambiguate and thus to improve the annotation precision on many levels. Most obviously, the parallel corpus helps to determine the correct word sense in a given sentence. For example, the word *Mönch* in our Alpine corpus may refer to a prominent mountain in central Switzerland or to a *monk* (= male

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

person in a monastery). If the corresponding sentence in the English or French translation also contains the word *Mönch*, then it is clear that the ambiguous word in the German sentence refers to the mountain name.

We have developed a similar kind of disambiguation for lemmas. For example, the German word form *gehört* may have the lemma *hören* (EN: to hear) or *gehören* (EN: to belong). Depending on the corresponding sentence in English or any other language, we can easily compute the correct lemma for the ambiguous word in the German sentence (Volk et al. 2016a). Of course, this kind of knowledge transfer between the languages in a parallel corpus presupposes word alignments across the languages.

2.3 Language-specific corpus annotation

In addition to these general considerations on annotation, many languages will have specific requirements. Compounding languages like German or Swedish will profit from compound segmentation in lemmatization. E.g. a German text might mention the compounds *Montblanc-Besteigung, Mont-Blanc-Expedition, Montblancgipfel* with or without a hyphen, and we will miss the mountain name *Montblanc* if we do not split these compounds and normalize the spelling variants. Splitting and normalization also facilitate cross-lingual word alignment since it reduces 1-to-many alignments.

In addition, splitting allows for the correct interpretation of elliptical compounds. Patterns like German *Schnee- und Eismassen* (EN: snow and ice masses) are an abbreviation for *Schneemassen und Eismassen*. The elliptical compound can be resolved by splitting the complete compound *Eis#massen* and adding the final element of the complete compound to the lemma of the elliptical compound (Aepli & Volk 2013).

Another example of a language-specific annotation is the re-attachment of German verb prefixes that occur separated in the sentence. In example sentence 2 the prefix *auf* (EN: on) needs to be re-attached to the verb stem *fällt* (EN: falls) in order to compute the correct verb lemma *auffallen* (EN: to strike, to notice) (Volk et al. 2016b).

(2) DE: Selber **fällt** mir der kleine Fehler aber kaum **auf**.
EN: However I do not notice the little mistake.

Our re-attachment algorithm is based on POS tags and re-attaches the separated prefix to the most recent preceding finite verb form when this results in a valid German prefix verb (from a manually curated list of about 8000 such verbs). It works with 96.8% precision when evaluated against manually re-attached prefixes in the TüBa/DZ treebank.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

For languages with large inflectional paradigms, for example agglutinative languages like Finnish, Turkish or Quechua, the word alignment on the level of word forms might not be practical. A noun form in Quechua, for instance, might correspond to a relative clause in English. In these cases it is advantageous to split the word form into its morphemes and to compute the alignment on the morpheme level. We have experimented with this kind of alignment between Spanish and Quechua, and it has resulted in a clear increase in precision and recall of the aligned items (Rios et al. 2012).

3. Aligning parallel corpora

Document alignment is the starting point of all alignment activities. If a corpus is built on OCRed text or on web-crawled text, then document alignment requires article boundary detection and subsequent document alignment based on properties such as author names, article lengths and publication dates.

The next step is sentence alignment, which is a precondition for any exploitation of parallel corpora. Sentence alignment can be computed efficiently based on length comparisons (based on character counts), co-occurring numbers, names and cognates. Hun-Align is a sentence aligner that implements a two-pass algorithm which does a rough alignment and builds a bilingual dictionary in the first pass and uses this dictionary in the second pass for refined sentence alignment. It works well for parallel texts that have corresponding sentences in the same order (monotonicity condition) and with few omissions.

BleuAlign (Sennrich & Volk 2011) is a tool for sentence alignment of noisy text. When aligning sentences between languages 1 and 2, BleuAlign uses a machine translation of the sentences from language 1 into language 2 and then compares the automatically translated sentences with the human-translated sentences in language 2 in order to determine the alignment. The comparison is done with the BLEU word-n-gram metric which was developed for machine translation evaluation.

Finally, we compute word alignments through GIZA, the Berkeley Aligner or FastAlign. Word alignment finds corresponding words or phrases in aligned sentence pairs. It can be computed on word forms or lemmas. Word alignment enables for example sorting the hits after translation variants (which may correspond to different word senses). It also enables annotation transfer (e.g. transferring name classes across languages) and cross-language disambiguation. Word alignment has opened many new avenues for linguistic research and translation studies over parallel corpora.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. Retrieval from parallel corpora

The Corpus Query Workbench has become a de facto standard for retrieval from annotated monolingual corpora. It allows simple and complex queries over words and their associated features (like POS tags, lemmas, name classes etc.). There is no such standard retrieval tool for parallel corpora.

Different commercial web sites offer searches over parallel corpora as a substitute or complement to bilingual dictionary searches. Most notably these are Linguee, Glosbe and Tradooit. But on these sites the texts do not have any linguistic annotation. SketchEngine is one of the few search systems that allows for query conditions to be specified over both sentences in a parallel sentence-aligned corpus. But it does not exploit automatically computed word alignment.

We are working towards such a flexible and powerful retrieval tool for parallel corpora. Our prototype system, Multilingwis,[1] allows for word form or lemma searches, for fixed sequence and bag-of-word searches and for the exclusion of function words from queries. We have also included the option to filter for source language. In this case, the query (in any of the supported languages) results only in hits with utterances that originate in a specific source language. For example, I could search for the German *Binnenmarkt* only in those cases where the original utterance is in French. This allows the researcher to distinguish between searches in original texts versus translations.

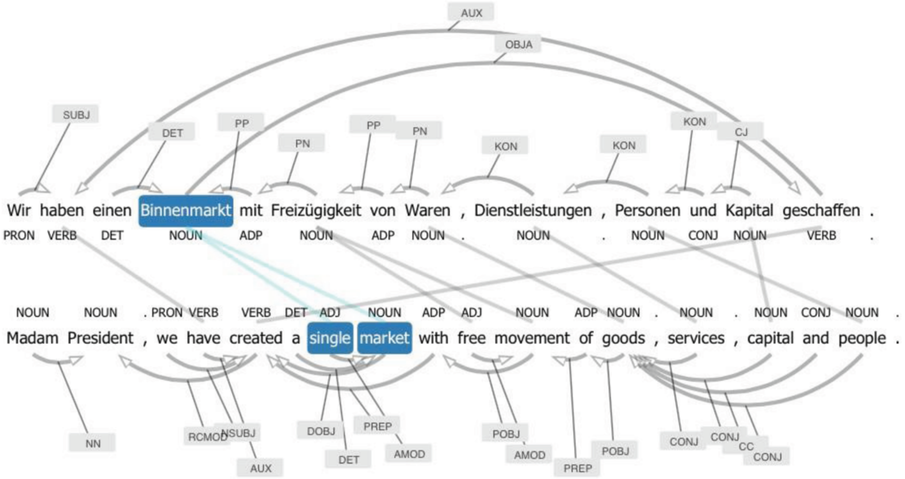

AUX
OBJA
SUBJ
DET
PP
PN
PP
PN
KON
KON
KON
CJ
Wir haben einen Binnenmarkt mit Freizügigkeit von Waren , Dienstleistungen , Personen und Kapital geschaffen .
PRON VERB DET NOUN ADP NOUN ADP NOUN . NOUN . NOUN CONJ NOUN VERB .
NOUN NOUN . PRON VERB VERB DET ADJ NOUN ADP ADJ NOUN ADP NOUN . NOUN . NOUN CONJ NOUN .
Madam President , we have created a single market with free movement of goods , services , capital and people .
NN
RCMOD
NSUBJ
AUX
DOBJ
DET
PREP
AMOD
POBJ
AMOD
PREP
POBJ
CONJ
CONJ
CC
CONJ

Figure 2. Query result for "single market" on Europarl German–English with full annotation display

[1] <pub.cl.uzh.ch/purl/multilingwis>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

In a second prototype, we have experimented with database searches over multi-parallel texts including displays of the hits as parallel dependency trees with POS tags and word alignment (see Figure 2).

Inspiration for designing retrieval systems for annotated and aligned parallel corpora may come from research on parallel treebanks. We have compiled small hand-crafted parallel treebanks and developed a tool with a powerful query language (Volk et al. 2011). A search system for larger parallel treebanks has been presented in (Meurer 2012).

Word alignment provides the basis for many different application scenarios. For example, word aligned corpora allow for translation discovery. We built a parallel corpus English–German of film and TV subtitles. When we queried for the German word *fragen* we found the obvious English translation variants *to ask, to say, to wonder, to question* (in this order of frequency). But next in the ranked list we found *to go,* which looked like an alignment error at first sight. But on closer inspection we discovered that this is a real translation option for German *fragen* as in the following example sentences.

(3) DE: … und sie **fragte** “Was ist das?”
EN: … and she **goes**, “What’s that?”

(4) DE: Ich war jung. Ich **fragte** “Wo ist England?”
EN: I was young. I **went**, “Wherés England?”

A special case of translation discovery is translation error detection. For example, we checked for translation variants of month names. When we queried for the English *July* we found that in about 1% of the translations it is erroneously translated with German *Juni*, French *Juin*, and Spanish *Junio* all meaning *June*. We observed this confusion in many of our parallel corpora. Obviously, the similarity of the month names *June* and *July* is confusing for human translators.

The process of translation error detection can be automated if word alignment works reliably and if the possible translations for a given query term are clearly defined (as is the case in the above example of month names).

Another application of parallel corpora is synonym detection through mirroring. Starting from a query in one language and then querying back from a translation hit in another language will lead to synonyms in the initial language. For example, when we searched for the idiomatic German adverb phrase *klipp und klar*, we found that *quite clear* and *very clearly* are among the top English translation variants in Europarl (cf. Figure 1). Now, if we query for *very clearly* in the opposite direction, we get the German *sehr deutlich, ganz klar* and *ganz eindeutig* which are synonymous expressions for the initial query phrase *klipp und klar*.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Queries over parallel corpora provide translation variant ranking based on corpus frequencies. These frequencies are obviously dependent on the corpus (more precisely, on the textual domain). For example, when querying for the German word *Leiter* (EN: leader or ladder or electrical conductor) we get different rankings for the French translation variants in our Alpine corpus in comparison to our corpus of Swiss laws. In the Alpine corpus the French *échelle* (EN: ladder) is ranked second after *chef* whereas in the Swiss Laws in French *échelle* is only on fifth place after *conducteur, directeur, chef, responsable*.

Augustinus et al. (2016) present a special user interface for example-based querying. This could be a possible future direction for large parallel corpora.

5. Conclusion

We have outlined a number of issues for the annotation and alignment of parallel corpora. We have argued that there are standard corpus annotation tools that work for many languages, and there are language-specific annotations that, for example, deal with elliptical compounds and separated verb prefixes in German. In addition, parallel corpora require automatic alignment not only on the sentence level but also for words and phrases. This level of alignment opens up many new applications such as translation discovery and translation error detection.

There is still no powerful and flexible search tool for annotated parallel corpora. We are developing Multilingwis, which already has some innovative features but still lacks a query language that comes close to the coverage of standard monolingual corpus query tools.

Acknowledgments

We would like to thank Johannes Graën for preparing the multiparallel Europarl corpus and Christof Bless for the visualizations of parallel parsed sentences. This research was supported by the Swiss National Science Foundation under grant 105215 146781 for “SPARCLING: Large Scale PARallel Corpora for LINGuistic Investigation” (2013–2017), a joint project with Marianne Hundt and Elena Callegaro at the English Department of the University of Zurich.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

References

Aepli, Noëmi & Volk, Martin. 2013. Reconstructing complete lemmas for incomplete German compounds. In *Proceedings of The International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, Irena Gurevych, Chris Biemann & Torsten Zesch (eds), 1–13. Darmstadt: Springer.

Augustinus, Liesbeth, Vandeghinste, Vincent & Vanallemeersch, Tom. 2016. Poly-GrETEL: Cross-lingual example-based querying of syntactic constructions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3549–3554. European Language Resources Association (ELRA).

Ebling, Sarah, Sennrich Rico, Klaper, David & Volk, Martin. 2011 Digging for names in the mountains: combined person name recognition and reference resolution for German alpine texts. In *Human Language Technology Challenges for Computer Science and Linguistics. LTC 2011* [Lecture Notes in Computer Science Vol. 8387], Zygmunt Vetulani, Joseph Mariani (eds), 189–200. Cham: Springer. doi: https://doi.org/10.1007/978-3-319-08958-4_16

Göhring, Anne & Volk, Martin. 2011. The Text + Berg corpus: An alpine French-German parallel resource. In *Proceedings of Traitement Automatique des Langues Naturelles* (*TALN* 2011), Montpellier, 27 Juni –1 Juli 2011.

Graën, Johannes, Batinic, Dolores & Volk, Martin. 2014. Cleaning the Europarl corpus for linguistic applications. In *Proceedings of KONVENS*, 222–227. Hildesheim.

Junczys-Dowmunt, Marcin, Pouliquen, Bruno & Mazenc, Christophe. 2016. Coppa v2.0: Corpus of parallel patent applications building large parallel corpora with gnu make. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora at LREC*, 15–19. Portorož, Slovenia.

Koehn, Philipp. 2005. Europarl: A parallel *corpus* for statistical machine translation. In *Proceedings of Machine Translation Summit X*, 79–86. Phuket.

Lison, Pierre & Tiedemann, Jörg. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 923–929. Portorož, Slovenia.

McDonald, Ryan & Nivre, Joakim. 2011. *Analyzing* and integrating dependency parsers. *Computational Linguistics* 37(1): 197–230. https://doi.org/10.1162/coli_a_00039

Meurer, Paul. 2012. INESS-Search: A search system for LFG (and other) treebanks. In *Proceedings of LFG12 Conference*, Miriam Butt & Tracy H. King (eds). Stanford, CA: CSLI Publications.

Petrov, Slav, Das, Dipanjan & McDonald, Ryan. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2089–2096. Istanbul.

Rios, Annette, Göhring, Anne & Volk, Martin. 2012. Parallel treebanking Spanish–Quechua: How and how well do they align? *Linguistic Issues in Language Technology* 7(1): 1–19.

Sennrich, Rico & Volk, Martin. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th International Nordic Conference of Computational Linguistics (Nodalida)*, 175–182. Riga.

Steinberger, Ralf, Pouliquen, Bruno, Widiger, Anna, Ignat, Carmelia, Erjavec, Tomaz, Tufis, Dan & Varga, Daniel. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20 + languages. In *Proceedings of* LREC, 2142–2147. Genoa.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Volk, Martin & Clematide, Simon. 2014. Detecting code-switching in a multilingual alpine heritage corpus. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 24–33, Doha, Qatar. https://doi.org/10.3115/v1/W14-3903

Volk, Martin, Graën, Johannes & Callegaro, Elena. 2014. Innovations in parallel corpus search tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 3172–3178. Reykjavik.

Volk, Martin, Amrhein, Chantal, Aepli, Noëmi, Müller, Mathias & Ströbel, Phillip. 2016a. Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of KONVENS*, 288–296. Bochum.

Volk, Martin, Clematide, Simon, Graën, Johannes & Ströbel, Phillip. 2016b. Bi-particle adverbs, PoS-tagging and the recognition of German separable prefix verbs. In *Proceedings of KONVENS*, 297–305. Bochum.

Volk, Martin, Marek, Torsten, & Yvonne, Samuelsson. 2011. Building and querying parallel treebanks. Translation: Computation, Corpora, Cognition (Special Issue on Parallel Corpora: Annotation, Exploitation and Evaluation) 1(1): 7–28.

Ziemski, Michał, Junczys-Dowmunt, Marcin & Pouliquen, Bruno. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 3530–3534. Portorož, Slovenia.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART II

Parallel corpora

Creation, annotation and access

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

InterCorp

A parallel corpus of 40 languages

Petr Čermák
Charles University Prague

This chapter presents the current version of InterCorp, a parallel corpus created at the Faculty of Arts, Charles University in Prague. The corpus contains texts in Czech aligned with one or more foreign-language version(s), including Czech and 39 other languages. The chapter analyses its structure and technical parameters, and describes some technical tools used with the corpus (Kontext, a corpus query interface, and InterText, a parallel text alignment editor created specifically for the project). Similarly, the contribution discusses Treq (Translation Equivalents Database), a collection of bilingual Czech-foreign language dictionaries built automatically from InterCorp. In the last section of the chapter, the possibilities for methodological and linguistic exploitation of the corpus are discussed.

Keywords: parallel corpus, InterCorp, comparison of languages, Spanish, Czech National Corpus

1. Introduction

The aim of this chapter is to present the current version of *InterCorp*, a parallel corpus created at the Faculty of Arts, Charles University in Prague. The original intention of the project, in 2005, was to create a parallel corpus for most of the languages studied at the Faculty of Arts. In a brief description of the first phase of the project (Čermák 2007), we wrote that we hoped the final corpus would contain textual data in 24 languages along with their Czech equivalents, and the Spanish part would comprise at least 4,500,000 lexical units. This chapter demonstrates, among other things, that *InterCorp* has radically surpassed our original expectations.

https://doi.org/10.1075/scl.90.06cer
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

2. Description of the corpus

InterCorp is a part of the vast *Czech National Corpus* (CNC), a group of various types of corpora that map the past and present of the Czech language with a great amount of linguistic data. This means that *InterCorp* was not built from scratch. On the one hand, there was a large quantity of existing linguistic data in Czech available for translation into other languages; on the other, a team of experienced specialists with sophisticated technical skills was available for work on the future parallel corpus.

It is only logical, therefore, that the pivot language of this corpus is Czech: The corpus contains texts in Czech (either in the original or in translation from other languages that form part of the corpus) aligned with one or more foreign-language versions (Czech and 39 other languages are included). This principle allows us to combine languages and to create a large number of bilingual, trilingual (etc.) subcorpora that logically differ in size; all "foreign" texts have Czech counterparts, whilst a foreign text does not necessarily have a counterpart in an additional foreign language. After registration, the use of *InterCorp* is free of charge because the aim of the project is to provide a source of data for theoretical studies, student research, foreign-language learning, etc., and to serve the general and academic public rather than generate profit.

The corpus can be accessed via a standard web browser from the integrated search interface *Kontext*. *Kontext* is a CNC-developed corpus query interface (developed by Tomáš Machálek, Machálek 2016), and it uses the Manatee server.

During the creation of *InterCorp*, various technical tools were used (for example, lemmatizers and taggers specific to different languages). The texts were aligned using *InterText*, a parallel text alignment editor created specifically for the *InterCorp* project but also used by other research groups (Vondřička 2016). This unique editor offers two different applications: InterText Server, a server application with a web-based interface for large projects, and InterText Editor, a desktop application enabling text alignment for personal use; for further details see Vondřička (2014).

Both of the CNC-developed tools, *Kontext* and *InterText*, are open-source tools available from <https://github.com/czcorpus>. The *InterCorp* project team encourages anyone who is interested to use them; in addition, assistance is offered with installation and configuration. Both tools are constantly being developed and supplemented with new functions, often based on user requests. *InterText*'s author, Pavel Vondřička, characterizes the tool as follows (Vondřička 2014: 1875):

> … the tool has been developed with flexibility in mind, and thus it can be used for several other purposes, including small personal projects or creation of translational databases. It has already been used in several other research projects world-wide, providing useful feedback.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

It is evident that the improvements made to the tools reflect two tendencies that are difficult to reconcile: a desire to make the corpus more perfect in technical terms, and the need also to make it more user-friendly.

> The goal is to further improve the usability, flexibility, scalability, and robustness (especially of the personal application InterText editor), in order to make it more and more user-friendly and useful both within the *InterCorp* project and for as many other similar purposes as possible, beyond the demands of the project itself. The tool is meant to generally help linguists and translation scholars (or translators) with various tasks related to alignment of texts, which are otherwise difficult without deeper technical skills. However, no application can offer both full flexibility and simplicity at the same time. Therefore, the user may often be in need to adapt various aspects of the software by means of configuring it for his own particular purpose. Improving configurability and a detailed documentation is thus also a priority. (Vondřička 2014: 1878)

As we can see, the *InterCorp* team makes every effort to ensure that not only the parallel corpus itself but also the technical expertise of its team members are of maximum benefit to the user.

The corpus comprises two basic parts, known respectively as the Core and the Collections. The Core consists mostly of fiction, and its texts were aligned manually by members of a team of nearly 200 individuals who cooperated on the collection and alignment of thousands of texts for this part of the parallel corpus. The intention is to have as many contemporary texts as possible, which means that, with a small number of exceptions, only texts originating after World War II have been used. Certain problems related to achieving a balanced mix of fiction-oriented texts (i.e. in the Core) are discussed in Rosen & Vavřín (2012).

The Collections offer automatically processed texts that, in the current version of the corpus, comprise:

– political commentaries published by Project Syndicate and VoxEurop (formerly PressEurop);
– a package of legal texts of the European Union from the Acquis Communautaire corpus;
– proceedings of the European Parliament dated 2007–2011 from the Europarl corpus; and
– film subtitles from the Open Subtitles database.

Users can also select a part of the corpus. If they use the Collections, they should expect a high number of misaligned segments due to the use of automatic alignment. On the other hand, the alignment of the Core texts is of a very high quality. In the current release of *InterCorp* (Version 9, Rosen & Vavřín 2016), the ratio between the

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Core and the Collections is approximately 1: 5.3 (the Core part: 231 million words in the aligned foreign language texts; the Collections part: 1,228 million words).

An updated version of *InterCorp* is released usually once a year. The current release consists of 1,460 million words in foreign languages, and 186 million words in Czech. (By comparison, Version 0 from 2008 comprised 25 million words in 19 foreign languages, and Version 1 from 2009 comprised 35 million words in 20 foreign languages.) Previous versions are still available (starting with Version 6).

The current version, Version 9, was published on September 9, 2016. It can be characterized as a referential, non-representative parallel corpus, consisting of Czech and 39 foreign languages. Table 1 offers the numbers of words in selected languages:

Table 1. Corpus size (in thousands of words)

Language	**Core**	**Collections**					**Total**
		Syndicate	**VoxEurop**	**Aquis**	**Europarl**	**Subtitles**	
Catalan	4,632	0	0	0	0	0	4,632
Dutch	11,444	314	2,955	24,746	15,563	29,362	84,386
English	21,208	3,818	2,670	24,207	15,580	52,101	119,586
French	12,406	4,393	2,928	27,351	17,178	25,961	90,219
German	31,168	3,725	2,482	21,723	13,089	8,366	80,556
Italian	8,694	651	2,707	24,849	15,489	14,653	67,046
Polish	21,433	0	2,378	20,627	12,811	26,572	83,822
Portuguese	2,605	369	2,999	28,602	16,484	43,391	94,454
Romanian	3,432	0	2,737	8,199	9,446	34,128	57.944
Russian	4,788	3,174	0	0	0	6,885	14,848
Spanish	19,310	4,324	2,816	27,001	15,885	36,378	105,716

Albanian, Arabic, Belarusian, Bulgarian, Croatian, Danish, Estonian, Finnish, Greek, Hebrew, Hindu, Icelandic, Japanese, Lithuanian, Latvian, Macedonian, Malay, Maltese, Norwegian, Romani, Serbian, Slovak, Slovenian, Swedish, Turkish, Ukrainian, and Vietnamese are also available.

As we can see in Table 1, the language subcorpora are not equal in terms of size. As a whole, the English subcorpus is the largest, while the German subcorpus offers the largest Core (i.e. the part that is manually aligned). Some languages are included only in the Core (i.e. Catalan), others only in the Collections (i.e. Vietnamese), which is why some language subcorpora (i.e. Arabic, Catalan, Hindu, Icelandic, Japanese, Malay, Romani, Vietnamese) are very small and in need of expansion.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

More than one half of the languages are lemmatized (21 out of 40) and a similar proportion include morphological annotation (24 out of 40). However, for each language, the selection of tools was autonomous and took a variety of aspects into account; the project team calls this approach *opportunistic*:

> The application of language-specific tools (tokenizers, morphological analyzers, taggers, lemmatizers) can be seen as an additional example of our opportunistic approach – all of them have been acquired ready-made, trained elsewhere on monolingual data using a language-specific tagset. Each of the language-specific tools may thus represent a different conceptual and practical solution to a number of issues: tokenization, lemmatization, patterning of word classes and morphological categories. While some of the decisions reflect real contrasts between individual languages, others show differences in theoretical backgrounds and formal approaches. (Rosen & Vavřín 2012: 2449)

2.1 The Spanish part of the corpus

The Spanish part consists of both Core and Collections, and is the second-largest language subcorpus of all. Its Core, the fifth-largest of all, comprises both Spanish and Hispanoamerican texts. (A list of the texts is available on the search site of the Spanish subcorpus under the *Choose texts* button.) The Spanish subcorpus

Table 2. Subcorpora size (in tokens; all texts are included in every languages, both originals and translations)

Corpus		Total	Core	Collections
Czech–Spanish	Czech part	114,209,832	20,537,645	93,672,187
	Spanish part	129,792,974	22,359,644	107,433,330
Spanish–English	Spanish part	105,880,681	5,920,572	99,960,109
	English part	108,108,209	6,099,523	102,008,686
Spanish–German	Spanish part	72,621,070	8,151,413	64,469,657
	German part	66,040,074	8,385,602	57,654,472
Spanish–French	Spanish part	87,211,273	3,780,536	83,430,737
	French part	89,442,739	4,119,726	85,323,013
Spanish–Italian	Spanish part	75,136,463	6,498,480	68,637,983
	Italian part	72,393,905	6,391,722	66,002,183
Spanish–Portuguese	Spanish part	95,516,329	1,503,255	94,013,074
	Portuguese part	100,305,816	1,659,799	98,646,017
Spanish–Catalan	Spanish part	4,541,983	4,541,983	0
	Catalan part	5,296,694	5,296,694	0

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

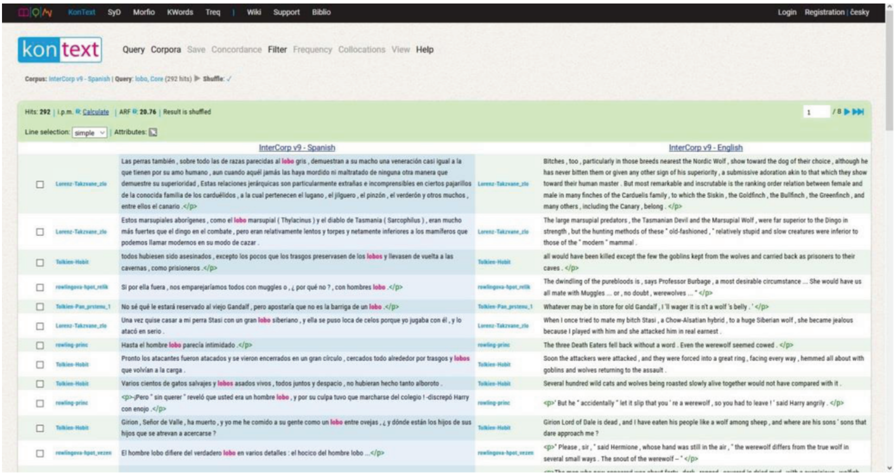
team pays specific attention to the development of bilingual subcorpora with English, German, and other Romance languages. Table 2 provides information as to the size of some of these subcorpora in tokens; we wish to emphasize that only the Core consists of fiction-oriented texts that enable selection of the original/target language and allow the researcher to take the direction of the translation into account:

The Spanish part of the corpus uses the *TreeTagger*, a tagger developed by Helmut Schmid at the Institute for Computational Linguistics of the University of Stuttgart (TreeTagger 2017).

3. Using the corpus

The search interface *Kontext* allows the user to compare up to four languages, selecting the source language and the aligned languages/subcorpora. Queries on one or more languages are possible by word form, by phrase, by CQL (Corpus Query Language) expression (including regular expressions) and by lemma and/or morphosyntactic tag (not in all languages, only in the lemmatized or tagged ones respectively). The interface also enables the further specification of the query according to context and to meta-information: the user can select the group of texts (Core, Collections), the source language, and the target language (in the Core part only), the particular text of a particular author (in the Core part only), and so on. The default setting for the query is the entire corpus.

The first piece of information offered by the interface on the query result (Figure 1) includes the number of hits, i.p.m. (instances per million positions;



Corpus: InterCorp v9 - Spanish | Query: lobo, Core (292 hits) ▶ Shuffle: ✓

Hits: 292 | i.p.m. Calculate | ARF: 20.76 | Result is shuffled

Line selection: simple | Attributes:

		InterCorp v9 - Spanish		InterCorp v9 - English
□	Lorenz-Takzvane_zlo	Las perras también , sobre todo las de razas parecidas al lobo gris , demuestran a su macho una veneración casi igual a la que tienen por su amo humano , aun cuando aquél jamás las haya mordido ni maltratado de ninguna otra manera que demuestre su superioridad , Estas relaciones jerárquicas son particularmente extrañas e incomprensibles en ciertos pajarillos de la conocida familia de los carduélidos , a la cual pertenecen el lugano , el jilguero , el pinzón , el verderón y otros muchos , entre ellos el canario .</p>	Lorenz-Takzvane_zlo	Bitches , too , particularly in those breeds nearest the Nordic Wolf , show toward the dog of their choice , although he has never bitten them or given any other sign of his superiority , a submissive adoration akin to that which they show toward their human master . But most remarkable and inscrutable is the ranking order relation between female and male in many finches of the Carduelis family , to which the Siskin , the Goldfinch , the Bullfinch , the Greenfinch , and many others , including the Canary , belong . </p>
□	Lorenz-Takzvane_zlo	Estos marsupiales aborígenes , como el lobo marsupial (Thylacinus) y el diablo de Tasmania (Sarcophilus) , eran mucho más fuertes que el dingo en el combate , pero eran relativamente lentos y torpes y netamente inferiores a los mamíferos que podemos llamar modernos en su modo de cazar .	Lorenz-Takzvane_zlo	The large marsupial predators , the Tasmanian Devil and the Marsupial Wolf , were far superior to the Dingo in strength , but the hunting methods of these " old-fashioned , " relatively stupid and slow creatures were inferior to those of the " modern " mammal .
□	Tolkien-Hobit	todos hubiesen sido asesinados , excepto los pocos que los trasgos preservasen de los lobos y llevasen de vuelta a las cavernas , como prisioneros .</p>	Tolkien-Hobit	all would have been killed except the few the goblins kept from the wolves and carried back as prisoners to their caves . </p>
□	rowlingova-hpot_relik	Si por ella fuera , nos emparejaríamos todos con muggles o , ¿ por qué no ? , con hombres lobo .</p>	rowlingova-hpot_relik	The dwindling of the purebloods is , says Professor Burbage , a most desirable circumstance ... She would have us all mate with Muggles ... or , no doubt , werewolves ... " </p>
□	Tolkien-Pan_prstenu_1	No sé qué le estará reservado al viejo Gandalf , pero apostaría que no es la barriga de un lobo .</p>	Tolkien-Pan_prstenu_1	Whatever may be in store for old Gandalf , I 'll wager it is n't a wolf 's belly . ' </p>
□	Lorenz-Takzvane_zlo	Una vez quise casar a mi perra Stasi con un gran lobo siberiano , y ella se puso loca de celos porque yo jugaba con él , y lo atacó en serio .	Lorenz-Takzvane_zlo	When I once tried to mate my bitch Stasi , a Chow-Alsatian hybrid , to a huge Siberian wolf , she became jealous because I played with him and she attacked him in real earnest .
□	rowling-princ	Hasta el hombre lobo parecía intimidado .</p>	rowling-princ	The three Death Eaters fell back without a word . Even the werewolf seemed cowed . </p>
□	Tolkien-Hobit	Pronto los atacantes fueron atacados y se vieron encerrados en un gran círculo , cercados todo alrededor por trasgos y lobos que volvían a la carga .	Tolkien-Hobit	Soon the attackers were attacked , and they were forced into a great ring , facing every way , hemmed all about with goblins and wolves returning to the assault .
□	Tolkien-Hobit	Varios cientos de gatos salvajes y lobos asados vivos , todos juntos y despacio , no hubieran hecho tanto alboroto .	Tolkien-Hobit	Several hundred wild cats and wolves being roasted slowly alive together would not have compared with it .
□	rowling-princ	<p>-¡Pero " sin querer " reveló que usted era un hombre lobo , y por su culpa tuvo que marcharse del colegio ! -discrepó Harry con enojo .</p>	rowling-princ	<p>' But he " accidentally " let it slip that you ' re a werewolf , so you had to leave ! ' said Harry angrily . </p>
□	Tolkien-Hobit	Girion , Señor de Valle , ha muerto , y yo me he comido a su gente como un lobo entre ovejas , ¿ y dónde están los hijos de sus hijos que se atrevan a acercarse ?	Tolkien-Hobit	Girion Lord of Dale is dead , and I have eaten his people like a wolf among sheep , and where are his sons ' sons that dare approach me ?
□	rowlingova-hpot_vezen	El hombre lobo difiere del verdadero lobo en varios detalles : el hocico del hombre lobo ...</p>	rowlingova-hpot_vezen	<p>" Please , sir , " said Hermione , whose hand was still in the air , " the werewolf differs from the true wolf in several small ways . The snout of the werewolf – " </p>

Figure 1. Query result offered by the interface (Spanish–English subcorpus, lemma *lobo*)

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

related to the subcorpus selected by the user), and ARF (Average Reduced Frequency). The results of the query can subsequently be used and processed in a number of ways. For example, the user can sort them according to different criteria, shuffle them to obtain a random sample, or apply additional contextual filters to further refine the results. At the same time, the interface offers complex statistical processing so that the user may acquire the frequency of lemmas, word forms etc. (also relative to the preset text types, which facilitates research on language variation) or search collocation candidates. He can also, of course, save the results in a format of his preference (xlsx, xml, txt, etc.).

4. Specific tools: Translation equivalents database

With more than 10 years of experience in developing this project, the *InterCorp* team is creating some specific, powerful supplementary instruments within the advanced parallel corpus. *Treq* (*Translation Equivalents Database*, <https://treq.korpus.cz/>) is a collection of bilingual, Czech–foreign language dictionaries built automatically using material from *InterCorp*. Its current version is based on texts from Release 9. In the creation of *Treq*, the first step was to align the original and the translated texts word-to-word using statistical methods provided by the *GIZA ++ program* (Och & Ney 2003). The aligned word pairs were then sorted and summarized. (For further information see Škrabal & Vavřín 2017.)

In other words, *Treq* constitutes a running phase of creating dictionaries for many languages, based on the data in the parallel corpus. The user chooses a source language and a target language (theoretically with all the 40 languages of *InterCorp* being available, although, as is clear from our description of the structure of the corpus, some language subcorpora are very small and limited in their ability to combine with other languages) and restricts the extension of the corpus (the default setting works with the entire corpus; however, it is also possible to work with the fiction-oriented Core texts only, or with specific Collections). The query can be entered as pertaining either to a specific word form or to a lemma. The latest version of *Treq* also offers a multiword query option in which the user can search for a combination of words, which constitutes a significant functional improvement in the database.

The query result is a list of all translation candidates for the given word. The term "translation candidates" is important: after automatic excerption and alignment the results were not reviewed, and so the relative frequency of a given pair may serve only as an approximate indicator of reliability. The user has to browse the list of translation candidates and consider the relevancy of the translation (click on a particular candidate, browse its occurrences in *InterCorp,* and check

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

the translation contexts). The list usually contains a few errors – there are translations that cannot be considered adequate – but in general, the database provides information that can be highly useful to a translator or lexicographer. In addition, further improvements to *Treq* are expected to be made in the near future.

5. Exploiting *InterCorp*

As mentioned above, the aim of the project is to offer a source of data for theoretical studies, student research, foreign-language learning, etc. The last ten years have proven beyond doubt that *InterCorp* really can be put to use in all of these areas.

The linguistic application of *InterCorp* runs along two basic lines that are intertwined and frequently inseparable:

1. the testing of the methodological possibilities of *InterCorp*; and
2. the comparison of languages using the data provided by *InterCorp*.

It comes as no surprise that, in the first phase of the existence of *InterCorp*, the majority of studies was created by researchers involved either directly in the project or in the field of Czech linguistics. The team of the *Czech National Corpus* as provider of the corpus organized various conferences in Prague dedicated to the creation and methodology of the corpus, paying special attention to the parallel corpus and *InterCorp*. The papers from those conferences (for example Čermák, Corness & Klégr 2010) recollect studies that test the parallel corpus *InterCorp*, exploring data from various languages included within it. Analyses of the findings of those studies and of user feedback have contributed not only to the continuous improvement of the corpus but also to synthetic theoretical reflection on the present state and the future of the project (Rosen & Vavřín 2012; Čermák & Rosen 2012; Rosen 2016; Nádvorníková 2017).

As the parallel corpus grows, the number of studies dedicated to the comparison of languages increases. Taking into consideration the fact that Czech is its pivot language and, as such, serves as the basis for all relations in the corpus, it is only logical and understandable that studies comparing Czech with another language predominate. Nevertheless, there have recently come into being many studies analysing two or more languages other than Czech, with only occasional reference to Czech as an "intermediate" language. To give an example from the field of Spanish/Romance linguistics, Štichauer and Čermák (2016) have undertaken a comparative analysis of causative constructions in Spanish, Italian, and Czech. In total there are hundreds of scientific works that have used the *InterCorp* to date, as is clearly demonstrated by the large quantity of texts in the repository of bibliographical items based on the Czech National Corpus (Repository 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

6. Conclusion

In this chapter, we have provided a brief description of one parallel corpus – *InterCorp* – that should be considered rather exceptional, for a number of reasons. Since 2005, the year of its initial release, the project has grown in many different ways: the number of languages included in the parallel corpus has significantly increased along with the size of the subcorpora for the individual languages; the tools and the technical base are under constant development, reflecting general progress in the field; the number of users has steadily grown; and many studies utilizing this methodology have been published, providing the project team with highly useful feedback. In short, *InterCorp* is now a thoroughly tested, fully functional tool that makes it possible to reliably compare 40 languages. You are invited to try it out for yourself.

Acknowledgment

The work was supported by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

References

Čermák, František, Corness, Patrick & Klégr, Aleš (eds). 2010. *InterCorp: Exploring a Multilingual Corpus*. Prague: Nakladatelství Lidové Noviny & Ústav Českého národního korpusu.

Čermák, František & Rosen, Alexandr. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 17(3): 411–427. https://doi.org/10.1075/ijcl.17.3.05cer

Čermák, Petr. 2007. Acerca de los corpora paralelos: El proyecto Intercorp (About the parallel corpora: The Intercorp project). *Verba* 34: 375–380.

Machálek, Tomáš. 2016. *Kontext*. <http://kontext.korpus.cz> (18 November 2017).

Nádvorníková, Olga. 2017. Pièges méthodologiques des corpus parallèles et comment les éviter (Methodological traps of parallel corpora and how to avoid them). *Corela. Cognition, Représentation, Langage* HS-21: 1–28.

Och, Franz Josef & Ney, Hermann. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1): 19–51. https://doi.org/10.1162/089120103321337421

Repository of bibliographical items based on the *Czech National Corpus*. 2017. <https://www.korpus.cz/biblio> (18 November 2017).

Rosen, Alexander & Vavřín, Martin. 2016. *Korpus InterCorp, version 9 of 9 Sep 2016*. Institute of the Czech National Corpus, Charles University, Prague 2014. <http://www.korpus.cz> (18 November 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Rosen, Alexandr & Vavřín, Martin. 2012. Building a multilingual parallel corpus for human users. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari et al. (eds), 2447–2452. Turkey: European Language Resources Association (ELRA).

Rosen, Alexandr. 2016. InterCorp – a look behind the façade of a parallel corpus. In *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*, Ewa Gruszczyńska & Agnieszka Leńko-Szymańska (eds.), 21–40. Warszawa: Instytut Lingwistyki Stosowanej.

Škrabal, Michal & Vavřín, Martin. 2017. The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, Kosek Iztok et alii (eds.), 124–137. Leiden: Lexical Computing CZ s. r. o.

Štichauer, Pavel & Čermák, Petr. 2016. Causative constructions of the *hacer / fare + verb* type in Spanish and Italian and their Czech counterparts: A parallel corpus-based study. *Linguistica Pragensia* 26(2): 7–20.

TreeTagger. 2017. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/> (18 November 2017).

Vondřička, Pavel. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Nicoletta Calzolari et al. (eds), 1875–1879. Reykjavik: European Language Resources Association (ELRA).

Vondřička, Pavel. 2016. *Intertext, Parallel Text Alignment Editor*. <http://wanthalf.saga.cz/intertext> (18 November 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Corpus PaGeS

A multifunctional resource for language learning, translation and cross-linguistic research

Irene Doval, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas and Barbara Lübke
University of Santiago de Compostela

This chapter presents the bilingual parallel corpus PaGeS, compiled by the research group SpatiAlEs from the University of Santiago de Compostela. PaGeS currently amounts to nearly 20 million tokens and consists of texts originally written in German and in Spanish and their correspondent translations into the other language, as well as a small portion of German and Spanish translations from third languages. The present contribution introduces the main characteristics of the PaGeS corpus, focusing on its design and compilation. It first explains the criteria for the selection of the texts and the details of text pre-processing, automatic alignment and manual review. It then addresses the search and display features describing the server architecture and indexing process. Finally, the intended development of the PaGeS corpus is briefly discussed.

Keywords: parallel corpora, corpus alignment, corpus visualization, Spanish, German

1. Introduction

The project[1] to create the bilingual parallel corpus PaGeS[2] (**Pa**rallel Corpus of **Ge**rman and **S**panish) emerged as a result of the research demands of linguists studying the expression of spatial relations and motion events in German and Spanish. The corpus is meant to provide reliable data of written language use as the basis for contrastive analysis in this field of investigation.

[1] This project (2014–2020) is being carried out at the University of Santiago de Compostela by the research team SpatiAlEs, led by Prof. Irene Doval.

[2] <http://corpuspages.eu>.

https://doi.org/10.1075/scl.90.07dov
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

In order to meet this purpose, PaGeS records samples of present-day German -including as well Swiss and Austrian texts- and Spanish -European and Latin American- from a variety of genres rich in lexical forms and grammatical patterns related to spatial notions. Source texts of both languages and their respective translations are carefully aligned, identifying corresponding sentences and smaller text segments in the original and translated versions. The query system allows monolingual as well as bilingual searches for words and sequences.

PaGeS has been conceived as a multi-modal and multi-functional language resource that will be publicly available. Among its multiple applications, it can be used in contrastive and general linguistics research in German and Spanish, translation studies, as well as in language teaching and language learning. Translators and learners at intermediate to advanced levels can use the corpus for finding multiple translation equivalents and contexts of use (Doval 2018: 182).

Existing multilingual corpora and other language resources that include German and Spanish bitexts offer valuable information on this pair of languages, but they scarcely assemble all the features required to meet the previously mentioned purposes. Some limitations relevant to our aims of research and application motivated our decision to build the PaGeS corpus.

First of all, most of the openly available multilingual resources are restricted to samples of highly specific text-types and domains, representing mainly legal and administrative or technical language use. This is the case of an important number of parallel corpora derived from source texts produced by the different institutions of the European Union (see Steinberger et al. 2014 for an overview). These are contained, for example, in the corpus included in *Multilingwis*[3] (Clematide et al. 2016), a search tool developed at the university of Zürich that covers the debates of the European Parliament in seven languages. The same applies to the major part of the large collection of parallel corpora included in the *Opus* project authored by Jörg Tiedemann at Uppsala University (Tiedemann 2009). Besides administrative documents from European institutions, *Opus*[4] covers journalistic texts and some minor collections from different online sources, as subtitles and technical documentation. Another huge multilingual parallel corpus is *InterCorp,*[5] compiled at Charles University in Prague (Čermák this volume). It covers 40 languages including German and Spanish and uses Czech as pivot language, that is all texts are aligned with a single Czech version (original or translation). A smaller part of the corpus, the so-called *core*, consists mostly of fiction, whereas

[3] <https://pub.cl.uzh.ch/projects/sparcling/multilingwis2.demo>.

[4] <http://opus.lingfil.uu.se/>.

[5] <http://ucnk.korpus.cz/intercorp/?lang=en>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

the greater part, the so-called *collections*, covers political commentaries, legal texts from EU-institutions, proceedings of the European Parliament and subtitles of movies and TV series.

Text-types specialized in legal, administrative or technical usage are not likely to provide the empirical data needed for linguistic inquiry into the expression of space and motion, our primary field of investigation, since they exhibit little variation in lexical forms and grammatical structures related to this semantic domain. Apart from linguistic research, the use of specialized corpora seems to be rather difficult in other fields of application, such as language teaching and language learning. As far as the fiction-oriented *core* part of *InterCorp* is concerned, users can search bilingual subcorpora for specific language pairs. In its current ninth release, the Spanish–German subcorpus contains 16.5 million tokens (Čermák, this volume). It has to be pointed out, however, that the major part of these are translations from third languages into Spanish and/or German, whereas direct translations represent only a very small part of them.[6] Other subcorpora German / Spanish included in multilingual parallel corpora only comprise fictional texts written in German and their translations into Spanish. That is the case of the German–Spanish subcorpus included in COVALT.[7] Still, this subcorpus is rather small, as it only contains 834,161 words (Molés-Cases & Oster, this volume).

Another very important factor to our research and to translation-based contrastive linguistics in general is the identification of source and target language. Investigation may be based on the analysis of direct translations between two languages under study, or may include the comparison of target texts translated from a third language as well. In any case, source and target texts have to be clearly differentiated and the direction of translation should be taken into account in order to derive sound conclusions from contrastive analysis. Available parallel corpora, however, do not always provide the relevant data. As far as documents of EU institutions are concerned, Steinberger et al. (2014: 6) observe:

> The source language for most documents produced by the EU institutions is no longer known. This information is not part of the explicit meta-information available for the documents. […] It is likely that at least some documents were translated via an intermediate language, that is that there are translations of translations.

[6] According to Rosen (2016: 29; 31), in the 8th release of InterCorp the Spanish–German subcorpus in the core had a size of 11 million words but included only 587 thousand words in Spanish translations from German originals, and 901 thousand words in German translations from Spanish originals.

[7] <http://www.covalt.uji.es>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

With respect to resources compiled automatically from bilingual web sites, such as *Linguee,*[8] it appears to be even more difficult to determine the source language of the included texts.

On the whole, we found that direct translations between German and Spanish are rather scarce or cannot be safely identified in available multilingual parallel language resources.

Through the compilation of PaGeS we hope to compensate for the problems of existing resources and provide an extensive base of empirical data suitable for contrastive research and bilingual information on German and Spanish. In the following sections, we present the characteristics of the corpus focusing on its composition (Section 2), the text preprocessing, (Section 3), the automatic alignment and its manual review (Section 4), the search and display features (Section 5) and the server architecture and indexing process (Section 6). Finally, Section 7 provides a brief summary and our outlook on the further development of the corpus.

2. Components and content

At the present stage (February 2018), the corpus contains 19.2 million tokens and 655.321 bisegments, that is pairs of aligned text chunks (sentences or smaller segments) from 104 sources. The German original texts and their translations into Spanish account for 46.1%, and the Spanish original texts together with their German translations make up 37.6% of the total tokens. A smaller part of the corpus (16%) contains translations into German and Spanish from other languages (so far English, French, Italian and Swedish). We think these texts provides useful data especially for translation studies since it allows the comparison of German and Spanish translations done from a third language source text. Table 1 gives a detailed account of numbers of tokens and sources at the present stage. Ongoing enlargement of the corpus is aimed at improving the balance of German and Spanish source language and at establishing a proportion of 10% of texts from a third language.

The compilation of the corpus has been restricted, for obvious reasons, to texts being available in both versions, original and direct translation, or, for a minor part of components, to being available in translation of a third-language source text into German and Spanish. For technical reasons, texts available in digital format were preferred. Apart from these conditions, the quality of the material has been a decisive criterion of text-selection. For this reason, only published editions,

[8] <http://www.linguee.com>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 1. Composition of PaGeS

Components	Sources	Tokens	Percent
German original text	54	4.253.900	22.4
German translation from Spanish	38	3.564.688	18.7
Spanish original text	38	3.584.908	18.9
Spanish translation from German	54	4.507.832	23.7
German translation from 3rd languages	12	1.577.794	8.0
Spanisch translation from 3rd languages	12	1.528.715	8.3
Total	**104**	**19.017.837**	

of original texts as well as translations, which had passed the control of established publishing houses, have been included. On the other hand, as mentioned before, we preferred text-types and genres that were expected to present a wide range of expressions related to spatial relations and motion events.

The vast majority of the corpus texts (95%) belong to samples of narrative fiction, dating, with very few exceptions, from the last 50 years (1967–2014); the bulk of texts are from the first decades of the current century. Although comprising a considerable range of subgenres from literary as well as from genre fiction, a significant part of the samples comes from crime, historical fiction, children's literature and young adult fiction, that is from genres particularly dense and varied in the description of spatial configurations and motion events. The corpus also covers a small proportion of non-fiction samples (5%) from genres like self-help books, travel diaries and biographies.

3. Text preprocessing, textual mark-up and metadata

After having selected the texts according to the criteria mentioned in Section 2, these have to be preprocessed and prepared for the alignment process. For this purpose, all texts are stored as text files in a txt-format, in atxt format, using the common character encoding UTF-8. The first step of preprocessing aims at reducing the noise and achieving as much parallelism as possible between source and target text in order to achieve better results with the alignment software. First of all, passages that do not belong to the body text are removed, that is, front matter, such as title, colophon, frontispiece, dedication, epigraph, contents, foreword, preface, and back matter, such as appendix, bibliography, author's and/or translator's notes, glossaries, etc. Moreover, chapters epigraphs, pictures and captions are discarded. Afterwards, the texts are carefully proofread in order to detect and amend errors

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

caused by the digitalization or conversion process. However, we do not correct spelling mistakes and do not adapt the German and Spanish original texts to the current spelling conventions.

Then, textual mark-up corresponding to the internal structure of the texts, that is the divisions in parts and chapters, is tagged. Smaller units, such as paragraphs, and further information concerning text formatting have not been considered yet.

Finally, descriptive metadata, that is "tags which encode descriptions of the corpus and its constituent texts" (Wynn 2008: 714–715) are collected and stored. The metadata list aims to be as detailed as possible, since we think this information could be extremely valuable for the conducting of linguistic and/or translational studies. For pairs of texts originally written in German and Spanish and then translated into Spanish and German, the metadata includes the following information: author's and translator's name, original and translated title, publication date and publisher of the original and translated version, original language and language version, and genre. Moreover, information about copyright, basic statistics (number of tokens, words and bisegments), and the name(s) of the alignment reviewer(s) are included. For texts originally written in another language and translated into German and Spanish, the metadata list additionally includes the original title and the publication year of the original version. This information is held in a separate database in tsv format and is linked to the single bitext files after the alignment process described in the following section.

4. Alignment[9]

Needless to say, the choice of an alignment tool constitutes one of the most fundamental decisions in the construction of a parallel corpus. In fact, the alignment and its accuracy play a crucial role for both the building and the exploitation of the corpus. For the corpus PaGeS, we decided to focus on sentence alignment systems, since this level of alignment is the most established for parallel corpora (see among others Tiedemann 2011: 37; Volk et al. 2014).

The choice of the alignment tool for PaGeS based on several tests conducted on five German and Spanish original texts and their corresponding translations

[9] The terminology used in this section is mostly based on Tiedemann (2011) and Zanettin (2012).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

using the following alignment tools: ABBYY Aligner,[10] bitext2tmx,[11] cwb-align,[12] LF-Aligner, based on Hunalign,[13] and Vanilla aligner.[14] LF Aligner achieved the highest alignment accuracy and was therefore chosen.

LF Aligner is a GUI wrapper that includes the Hunalign sentence alignment system (Varga 2012: 199). Hunalign combines both a *length-based* and a *lexical matching* approach and is therefore a so-called hybrid algorithm (Tóth et al. 2008; Varga et al. 2005). The alignment process runs in three main steps:[15] (1) Hunalign "builds alignments using a simple similarity measure" (Steinberger et al. 2006: 5), which is based "on sentence lengths and the ratio of identical words" (Steinberger et al. 2006: 5); (2) it builds a bilingual lexicon based on this first alignment; (3) it returns the alignment also taking into consideration the lexical similarity by means of the created dictionary (see Varga et al. 2005; Steinberger et al. 2006 and Varga 2012).

The alignment accuracy mainly depends on the degree of correspondence between source and target text. Obviously, a one-to-one correspondence is not always possible since during the translation process sentences can or have to be split, merged or be reordered and the translator may choose to insert or omit sentences or whole text passages (see among others Tiedeman 2011: 9; Varga 2012: 94). The genre also plays a very important role and in this regard, the alignment of literary texts may be more difficult than technical ones (Zanettin 2012: 155). Moreover, within literary texts, the degree of correspondence varies depending on the author, the translator, the texts themselves and on the direction of translation. Furthermore, as mentioned in Section 2, PaGeS also includes bitexts in which both the German and the Spanish bitext halves are translations from a third language and these "are particularly challenging in this aspect [alignment], since they have undergone two independent translation processes" (Doval 2016: 93).

[10] Available on <https://www.abbyy.com/en-eu/aligner/>. The PELCRA Polish-Russian parallel corpus was aligned with ABBY Aligner (Łaziński & Kuratczyk 2016).

[11] Available under <http://bitext2tmx.sourceforge.net/>.

[12] Included in the IMS CWB Open Corpus Workbench (Evert & CWB Development Team 2016).

[13] LF Aligner is available under <https://sourceforge.net/projects/aligner/>. *Hunalign* was used, for example, for the alignment of Intercorp (Čermák & Rosen 2012) and in the platform Multilingwis (Clematide et al. 2016). It is also part of the Uplug-tool used for the building of the OPUS corpora.

[14] Described in Danielsson & Ridings (1997) and used for the alignment of the Dutch Parallel Corpus (Macken et al. 2007).

[15] For a detailed description of the algorithm and of how it runs if a bilingual lexicon is provided from the beginning, see Varga (2012: 92–119).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

The inclusion of the aligned bitexts in the corpus PaGeS is based on two critieria. The first one is the percentage of empty alignments, for instance, a segment in one bitext half that has been linked to an empty segment in the other bitext half ($S_{src} > \varnothing_{trg}$ or $\varnothing_{src} > S_{trg}$). The second one is the percentage of segments containing more than 350 characters.[16] We generally discard bitexts showing more than 10% of empty alignments and/or 20% of segments longer than 350 characters, since their manual review and processing would be too time-consuming. However, we also take into consideration the location of the empty alignments and the long segments in the bitexts. If they are concentrated in concrete passages (e.g. long passages omitted, added or free translated), these are eliminated in both the source and the target bitext halves and the rest of the bitext is included in the corpus. Up until now, 17 of 134 bitexts have been definitively excluded. Further, we developed an effective procedure to manually review and validate the results of the alignment of the selected bitexts, and hence to improve its quality. For this purpose, we export them in Excel or Google Spreadsheets. We then count the number of characters of each segment separately and calculate the ratio for each bisegment.

After splitting those segments longer than 350 characters, we identify empty alignments, which are of two types in our bitexts. On the one hand, a segment in the source bitext half may have no correspondence in the other half because of an omission or an addition. Table 2 and 3, respectively, illustrate these cases of one-to-zero correspondence and zero-to-one correspondence:

Table 2. Omission of a segment in the target text

German source text	Spanish target text
Und es wäre eine Erklärung dafür, dass Goldbergs Sohn keine vierundzwanzig Stunden, nachdem wir die Leiche seines Vaters gefunden haben, mit einer ganzen Streitmacht auftaucht, um uns an weiteren Ermittlungen zu hindern.	Es más, eso explicaría también que, después de que encontráramos el cadáver de su padre, el hijo no haya tardado ni veinticuatro horas en presentarse aquí con las fuerzas armadas al completo para impedirnos seguir con las investigaciones.
Entweder Goldberg junior oder jemand anderes hat beste Beziehungen und ein Interesse daran, die sterblichen Überreste seines Vaters so schnell wie möglich verschwinden zu lassen.	
Goldbergs Geheimnis sollte geheim bleiben.	El secreto de Goldberg tenía que seguir oculto.

[16] We think that segments longer than 350 characters would hinder the results' visualization and the identification of the equivalent of a searched term in the other bitext half.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 3. Addition of a segment in the target text

Spanish source text	German target text
Llegó septiembre.	Es wurde September.
	Mit Arnau ging es allmählich aufwärts.
Bernat ya había visto sonreír y gatear a su hijo por la cueva y sus alrededores.	Bernat hatte seinen Sohn bereits lächeln sehen und er machte auf allen vieren Ausflüge durch die Höhle und in die nähere Umgebung.

In these cases, the empty segment is tagged according to the translation direction. Empty segments resulting from an omission in the target text are tagged as `n_t_s` (i.e. *not translated segment*) and those resulting from an addition are tagged as `a_s_t` (i.e. *added segment in translation*). In the case of bitexts derived from third languages, the empty segment is tagged as […] since both the German and the Spanish bitext halves are translations and it is impossible to determine whether the empty segment is of type `n_t_s` or `a_s_t`.

On the other hand, an empty alignment is not always the result of a one-to-zero or zero-to-one correspondence and can be due to a misalignment, for instance, a segment in one bitext half that the aligner did not match to its correspondence in the other half. These misalignments are generally occasioned by one-to-two or two-to-one correspondences, as shown in Table 4 and 5, respectively.

Table 4. Empty alignment due to a misalignment (one-to-two correspondence)

La azafata le rozó para desensimismarle.	Die Stewardess tippte ihm auf die Schulter und riß ihn aus seiner Versunkenheit.
	Sie deutete mit einem Lächeln ihres vollen, gesunden Gesichtes auf den Sicherheitsgurt.
Le indicó el cinturón con una sonrisa llena de carne sana y rouge enmarcada por una cabellera castaña, casi pelirroja, de las que no se encuentran en España.	Sie hatte Rouge auf den Wangen, und ihr langes Haar war von einem Kastanienbraun, das ins Rötliche spielte, wie man es in Spanien nicht findet.

Table 5. Empty alignment due to a misalignment (two-to-one correspondence)

Man werde sich einigen, sagte der Herzog.	Llegarían a un acuerdo, dijo el duque.
Ein Professorentitel sei möglich.	
Wenn auch nicht bei doppelten Bezügen.	La cátedra era posible, aunque sin doble sueldo.

After reviewing empty alignments, we focus on bisegments that despite having been paired do not correspond to each other or at least partially, as shown in Table 6.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 6. Empty alignment due to a misalignment

Saladin hat dem Tempelritter das Leben geschenkt.	Saladino le perdonó la vida al templario, fue un buen acto, un mizwa, como tú lo llamas, que propició otro, pues al templario debes agradecerle que tu querida hija siga con vida, no lo olvides.
Das war eine gute Tat, eine Mizwa, wie du es nennst, die sogleich eine andere Mizwa nach sich gezogen hat, denn diesem Tempelritter hast du zu verdanken, dass deine geliebte Tochter noch lebt, vergiss das nicht.« Recha öffnete den Mund, um ihr zu widersprechen, doch Daja legte ihr begütigend die Hand auf den verbundenen Arm und das Mädchen presste die Lippen zusammen und schwieg.	Recha abrió la boca para replicar, pero Daja le puso una mano tranquilizadora en el brazo vendado y la niña apretó los labios y calló.

To identify them, we look at the ratio and focus on those bisegments with a value between 7 and – 7, since we discovered that these tend to display misalignment and have therefore to be checked and realigned as necessary. Beside the degree of correspondence between bitext halves, segmentation is another aspect that should be taken into consideration, as pointed out by Tiedemann (2011: 9-11). We are currently examining to what extent improvements in the texts' automatic splitting could improve the alignment results. For instance, the sentence splitter does not consider some punctuation marks as sentence boundaries, such as "…" both in German and Spanish, "?" and "!" before "–", and ";" in Spanish, as shown in Table 7.

Table 7. Automatic splitting of sentences containing … and? –

Segmentation of the German text	Segmentation of the Spanish text
Man wollte eine Papierfabrik in Krakau bauen, die Metallindustrie in Riga auf Vordermann bringen, eine Zementfabrik in Tallinn errichten und so weiter.	Se abrió una fábrica papelera en Cracovia, se reformó una industria metalúrgica en Riga, una fábrica de cemento en Tallin … **La dirección del CADI, compuesta por pesos pesados del mundo de la banca y de la industria suecas, repartió el dinero.**
Die Gelder wurden von den Vorständen des SIB verteilt, lauter einflussreichen Persönlichkeiten aus der Welt der Banken und der Großindustrie.«	–¿Te refieres al dinero de los contribuyentes**? – Alrededor del cincuenta por ciento provenía de subvenciones estatales; el resto lo pusieron los bancos y la industria.**
»Steuergelder also?«	
»Ungefähr 50% staatliche Zuschüsse, den Rest steuerten die Banken und die Unternehmen selbst bei.	

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Introducing a boundary before the two sentences marked in bold made in this case a difference in the alignment results, as shown in Table 8 and 9, respectively:

Table 8. Alignment based on the original sentence splitting

German target text	Spanish target text
Man wollte eine Papierfabrik in Krakau bauen, die Metallindustrie in Riga auf Vordermann bringen, eine Zementfabrik in Tallinn errichten und so weiter.	Se abrió una fábrica papelera en Cracovia, se reformó una industria metalúrgica en Riga, una fábrica de cemento en Tallin … La dirección del CADI, compuesta por pesos pesados del mundo de la banca y de la industria suecas, repartió el dinero.
Die Gelder wurden von den Vorständen des SIB verteilt, lauter einflussreichen Persönlichkeiten aus der Welt der Banken und der Großindustrie.« »Steuergelder also?«	–¿Te refieres al dinero de los contribuyentes? –Alrededor del cincuenta por ciento provenía de subvenciones estatales; el resto lo pusieron los bancos y la industria.
»Ungefähr 50% staatliche Zuschüsse, den Rest steuerten die Banken und die Unternehmen selbst bei.	

Table 9. Alignment based on the improved sentence splitting

German target text	Spanish target text
Man wollte eine Papierfabrik in Krakau bauen, die Metallindustrie in Riga auf Vordermann bringen, eine Zementfabrik in Tallinn errichten und so weiter.	Se abrió una fábrica papelera en Cracovia, se reformó una industria metalúrgica en Riga, una fábrica de cemento en Tallin …
Die Gelder wurden von den Vorständen des SIB verteilt, lauter einflussreichen Persönlichkeiten aus der Welt der Banken und der Großindustrie.«	La dirección del CADI, compuesta por pesos pesados del mundo de la banca y de la industria suecas, repartió el dinero.
»Steuergelder also?«	–¿Te refieres al dinero de los contribuyentes?

These types of differences led to a hypothesis that considering more punctuation marks as sentence-boundary markers and exploring systematic differences in the German and Spanish punctuation systems could imply improvements in the alignment accuracy.

Overall, although being rather laborious and time-consuming, the procedures described in this section ensure a very high degree of alignment accuracy, which is essential for our research purposes. Moreover, and crucially, the manual reviewing process will also contribute to improving alignment accuracy of new bitexts, since the manually validated bitexts will be incorporated into a training corpus by means of which the aligner tool will be regularly retrained.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

5. Search and display features

As Dörk and Knight (2015: 84) assess, many existing corpora are "aimed mainly at people with expertise with linguistics". Thus, there is less reflection on the decisions involved in designing search and visualization of corpora and corpus analysis for non-expert users. As mentioned at the beginning of this chapter, we do not want to limit the target users of the corpus PaGeS to expert ones. Therefore, and in order to make the corpus a really multipurpose tool, that is useful for very different user groups, from cross-linguistics and translation researchers to lexicographers and NLP researchers to occasional or regular users, as well as German or Spanish learners, it is essential to provide an adequate interface for displaying and retrieving data, including corpus texts, metadata and linguistic annotation. To achieve this aim the corpus should provide the following functionalities (Doval 2018: 191ff):

a. *Fast search*: Given that a significant increase in size of the corpus PaGeS is planned, the search engine must allow searches in a quick and efficient way through large amounts of language data.
b. *User-friendly search*: The query language must be as simple as possible. An advanced, more complex, query language is only displayed if required. In addition, the search habits of users on the internet should be exploited, and thus, the query language of Google should serve as a model.
c. *Multi-level search*: The search system must allow queries across multiple layers of linguistic annotation, such as lemmatization and POS tagging.
d. *Display*: The search results must be displayed in an easy-to-read format. The matching segments have to be displayed side-by-side, and both the search word or phrase and its potential equivalent have to be highlighted.

To match the requirements of the above-mentioned users we have designed a three-level search. The first one, whose provisional web interface is shown in Figure 1, is the simple or standard search. In this case, the user only has to enter in the search field the search term (a word or a phrase) in German or Spanish. In these types of queries, lemmatization is applied by default. With multiword queries, all search words within a specific distance are found. Similar to Google search, if the term is enclosed between quotation marks " ", the search only returns results that exactly match the entered word form or phrase.

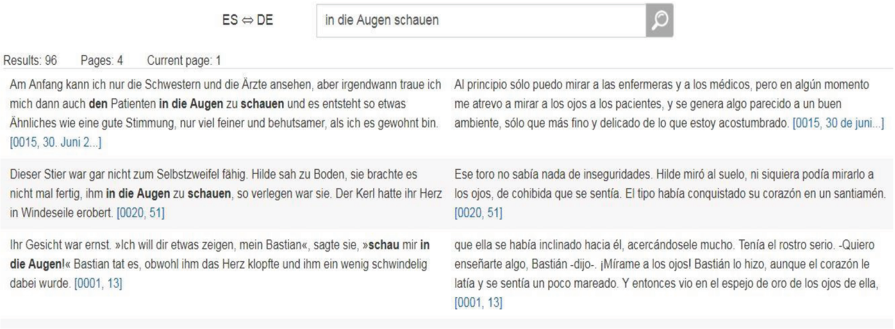
As Wynne (2008: 706) points out, "the most popular way to display the results of a search in a corpus is in the form of a concordance" and the most common concordance format is the KWIC (Key Word in Context) concordance, that is the node word is in a central position with all lines vertically aligned around the node. This presentation of the results, properly sorted, is very useful in monolingual

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

corpora for visualizing patterns of use. However, in bilingual corpora, the KWIC format cannot be considered user-friendly, since one of the main applications of these corpora is to quickly find possible equivalents of a search term.

For this reason, we decided for the visualization of the query results in a two-column html table, where one column corresponds to source texts and the other one to target ones. The search term is displayed in a cell with some context and is highlighted in bold. Depending on whether the search term is located in a source or a target text, it is shown in one column or in the other.

Occurrences in source texts are displayed in the left column, while those in target texts are shown in the right one. The correspondent segment is displayed in the same row but in a cell of the other column. With each query, information on the number of hits, the total number of pages, as well as the current page number is shown. The number of segments per page is fixed at 30, but this could be easily changed in the web application configuration file. Figure 1 provides an example of the standard search menu and some of its features:



ES ⇔ DE

in die Augen schauen

Results: 96 Pages: 4 Current page: 1

Am Anfang kann ich nur die Schwestern und die Ärzte ansehen, aber irgendwann traue ich mich dann auch **den** Patienten **in die Augen** zu **schauen** und es entsteht so etwas Ähnliches wie eine gute Stimmung, nur viel feiner und behutsamer, als ich es gewohnt bin. [0015, 30. Juni 2...]	Al principio sólo puedo mirar a las enfermeras y a los médicos, pero en algún momento me atrevo a mirar a los ojos a los pacientes, y se genera algo parecido a un buen ambiente, sólo que más fino y delicado de lo que estoy acostumbrado. [0015, 30 de juni...]
Dieser Stier war gar nicht zum Selbstzweifel fähig. Hilde sah zu Boden, sie brachte es nicht mal fertig, ihm **in die Augen** zu **schauen**, so verlegen war sie. Der Kerl hatte ihr Herz in Windeseile erobert. [0020, 51]	Ese toro no sabía nada de inseguridades. Hilde miró al suelo, ni siquiera podía mirarlo a los ojos, de cohibida que se sentía. El tipo había conquistado su corazón en un santiamén. [0020, 51]
Ihr Gesicht war ernst. »Ich will dir etwas zeigen, mein Bastian«, sagte sie, »**schau** mir **in die Augen**!« Bastian tat es, obwohl ihm das Herz klopfte und ihm ein wenig schwindelig dabei wurde. [0001, 13]	que ella se había inclinado hacia él, acercándosele mucho. Tenía el rostro serio. -Quiero enseñarte algo, Bastián -dijo-. ¡Mírame a los ojos! Bastián lo hizo, aunque el corazón le latía y se sentía un poco mareado. Y entonces vio en el espejo de oro de los ojos de ella, [0001, 13]

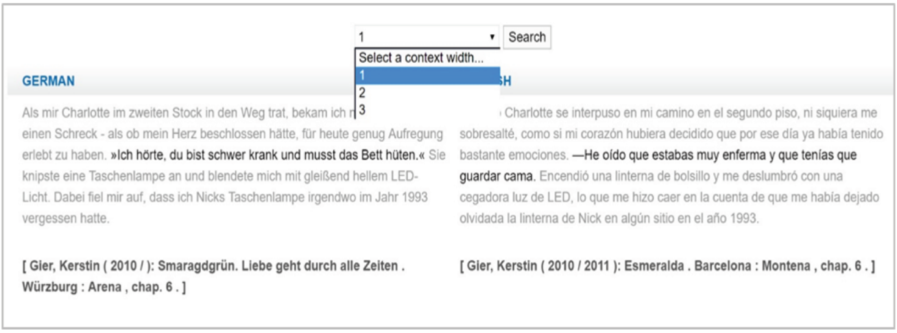
Figure 1. Standard search menu

At the bottom of the table, a set of links are available to allow the user to navigate through the pages and to download the query results in three formats: Excel, ODS and CSV (only available for registered user).

In each cell, information concerning the text ID, the corresponding part or chapter name and/or number are displayed in blue square brackets. By clicking on the work ID, the user can see a larger linguistic context and select the number of segments above and/or below they will see. Moreover, this screen shows detailed information concerning the bibliographic information, as shown in Figure 2.

The second level of search, as of yet not implemented, is in the advanced one. At this level, the user can control and restrict the scope of the search by applying the following drop-down search filters: text ID, author, publication year, original or translated text, translation from third languages and part-of-speech-tags. The occurrences are expected to be sorted by various criteria.

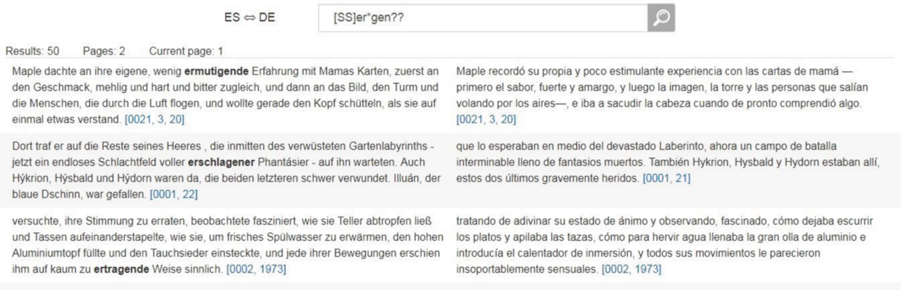
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



GERMAN	H
Als mir Charlotte im zweiten Stock in den Weg trat, bekam ich n einen Schreck - als ob mein Herz beschlossen hätte, für heute genug Aufregung erlebt zu haben. **»Ich hörte, du bist schwer krank und musst das Bett hüten.«** Sie knipste eine Taschenlampe an und blendete mich mit gleißend hellem LED-Licht. Dabei fiel mir auf, dass ich Nicks Taschenlampe irgendwo im Jahr 1993 vergessen hatte.	Charlotte se interpuso en mi camino en el segundo piso, ni siquiera me sobresalté, como si mi corazón hubiera decidido que por ese día ya había tenido bastante emociones. **—He oído que estabas muy enferma y que tenías que guardar cama.** Encendió una linterna de bolsillo y me deslumbró con una cegadora luz de LED, lo que me hizo caer en la cuenta de que me había dejado olvidada la linterna de Nick en algún sitio en el año 1993.
[Gier, Kerstin (2010 /): Smaragdgrün. Liebe geht durch alle Zeiten . Würzburg : Arena , chap. 6 .]	**[Gier, Kerstin (2010 / 2011): Esmeralda . Barcelona : Montena , chap. 6 .]**

Figure 2. Scope of context and bibliographic information

The last search level is the most complex one, and is currently partially available over the standard search interface. This level gives the user full command of the query syntax used in the underlying query tool: Solr. (Version 5.0.0).[17] This supports searches using regular expressions (RegEx).[18] The search term has to be preceded by [SS] (Solr Search). The search expression has to be constructed word by word, and for each word, it will be possible to specify several parameters. Figure 3 shows a formal search.



Results: 50 Pages: 2 Current page: 1

Maple dachte an ihre eigene, wenig **ermutigende** Erfahrung mit Mamas Karten, zuerst an den Geschmack, mehlig und hart und bitter zugleich, und dann an das Bild, den Turm und die Menschen, die durch die Luft flogen, und wollte gerade den Kopf schütteln, als sie auf einmal etwas verstand. [0021, 3, 20]	Maple recordó su propia y poco estimulante experiencia con las cartas de mamá —primero el sabor, fuerte y amargo, y luego la imagen, la torre y las personas que salían volando por los aires—, e iba a sacudir la cabeza cuando de pronto comprendió algo. [0021, 3, 20]
Dort traf er auf die Reste seines Heeres , die inmitten des verwüsteten Gartenlabyrinths - jetzt ein endloses Schlachtfeld voller **erschlagener** Phantásier - auf ihn warteten. Auch Hýkrion, Hýsbald und Hýdorn waren da, die beiden letzteren schwer verwundet. Illuán, der blaue Dschinn, war gefallen. [0001, 22]	que lo esperaban en medio del devastado Laberinto, ahora un campo de batalla interminable lleno de fantasios muertos. También Hykrion, Hysbald y Hydorn estaban allí, estos dos últimos gravemente heridos. [0001, 21]
versuchte, ihre Stimmung zu erraten, beobachtete fasziniert, wie sie Teller abtropfen ließ und Tassen aufeinanderstapelte, wie sie, um frisches Spülwasser zu erwärmen, den hohen Aluminiumtopf füllte und den Tauchsieder einsteckte, und jede ihrer Bewegungen erschien ihm auf kaum zu **ertragende** Weise sinnlich. [0002, 1973]	tratando de adivinar su estado de ánimo y observando, fascinado, cómo dejaba escurrir los platos y apilaba las tazas, cómo para hervir agua llenaba la gran olla de aluminio e introducía el calentador de inmersión, y todos sus movimientos le parecieron insoportablemente sensuales. [0002, 1973]

Figure 3. Formal search menu

Finally, this level will allow the combination search of words or phrases and POS tags (see Section 7). This formal search is naturally very powerful in executing precise search queries, but not particularly user-friendly. It is intended for more demanding users, such as researchers in contrastive linguistics or translation, who usually need a very specific subset of results, only possible with complex queries including a large set of parameters.

[17] <http://lucene.apache.org/solr/> (12 February 2018).

[18] <https://www.regular-expressions.info/> (12 February 2018).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

6. Server architecture and publishing data

The PaGeS corpus platform is mainly composed by a search engine and a web application. Both components are installed in an Ubuntu virtual machine accessible via a web interface from the internet. The search engine is a Solr server that contains all corpus indexed data and a lemma dictionary for the lemmatized search. The web application is developed with java Grails framework[19] and deployed in a Tomcat[20] server. This web application contains a component including Java methods, that calls the *solr-solrj* library.[21] This library is a Solr java client that communicates the PaGeS Web Application with the Solr server.

After the alignment review described in Section 4, some processes have to be performed in order to publish in a Solr server, all the information of the aligned bitexts. The result of the alignment manual review is a set of text files with the following tsv-format:

```
Text_Segment_ID<TAB>Text_Segment_Lang_1<TAB>Text_Segment_Lang_2
```

First, it is necessary to generate a lemma dictionary for each language in order to perform the lemmatized search.[22] This kind of dictionary consists of sets of word clusters, whereas each cluster groups all words appearing in the corpus under the same lemma. The lemma information was obtained through TreeTagger[23] for German and FreeLing[24] for Spanish, the tools we have chosen for POS tagging after several performance tests.[25] Both dictionaries are stored in a configuration file of the Solr server and are used for query expansion at searching time. Each dictionary contains a comma-separated word cluster per line, according to the following format:

[19] <https://grails.org/> (12 February 2018).

[20] <http://tomcat.apache.org/> (12 February 2018).

[21] <https://wiki.apache.org/solr/Solrj> (29 August 2017).

[22] The lemmatizer of the TreeTagger is not able to correctly lemmatize verbs when they occur with verb stem and particle split. Volk et al. (2016) have developped an algorithm to re-attach the separated prefix to the verb. This script was kindly made available to the corpus PaGeS by Prof. Volk and it is intended to implement it soon.

[23] <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/> (29 August 2017).

[24] <http://nlp.lsi.upc.edu/freeling/node/1> (29 August 2017).

[25] However, in the future we intend to test again the TreeTager with the Spanish texts after having trained it on a hand tagged subcorpus of our texts. The mentioned tests were carried out with the standard parameter files distributed with TreeTagger, without any additional manually training corpus (see Doval 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

```
word_1, word_2, …
…
```

The next step consists of adding the metadata mentioned in Section 3. This information is actually stored in a text file (in "field=value" format). Each reviewed bitext (stored as.tsv file) has a corresponding metadata document. Finally, all information (aligned segments and metadata) is indexed at the Solr search engine. Solr provides a fast indexing and quick searching tool for the corpus. As Solr is a general purpose search engine, data and information can be respectively structured and encoded without many restrictions.

In order to deal with this goal, a Tsv2Solr java program was implemented. This software calls the above-mentioned solr-solrj library again and then reads the content of tsv aligned text files and metadata files, opens a new Solr server session, eventually deletes all previous information and adds the new data. Figure 4 shows the PaGeS architecture and both the components and technologies used.

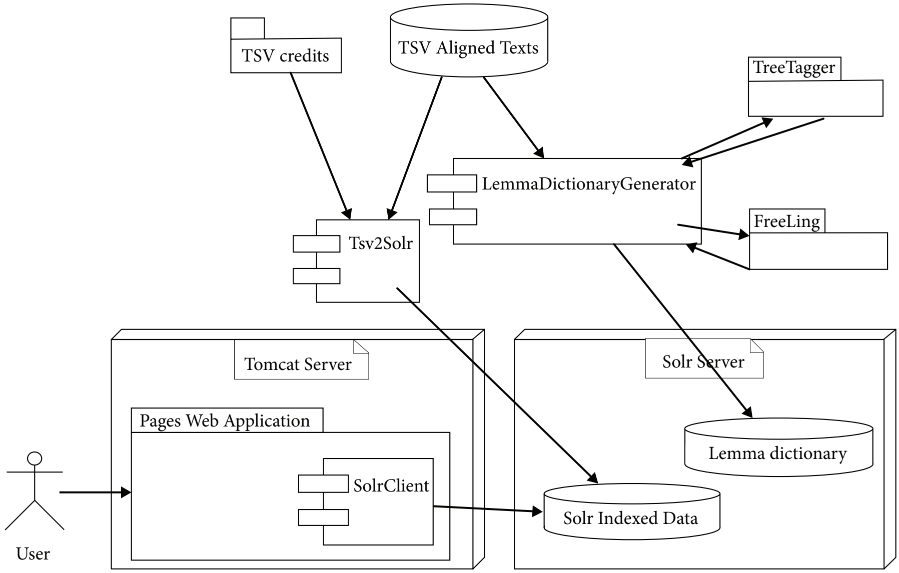


Figure 4. PaGeS platform architecture

7. Summary and outlook

In the previous sections, we have presented the main characteristics of the corpus PaGeS and the processes involved in its compilation and its indexation. Starting with very specific research objectives (Section 1), we have created a valuable parallel corpus German/Spanish, which includes texts carefully selected (Section 2) and

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

offers a very high alignment quality (Section 4). The three type of search levels (Section 5) are directed to very different users and thus will make the corpus PaGeS a multifunctional resource with an enormous potential. Its concrete applications in contrastive linguistics and language learning are currently being exploited within our research group (Doval 2018; Lübke & Liste Lamas 2019).

We are of course aware of some of the limitations of our corpus and intend to introduce new functionalities as soon as possible. As reported in Section 5, we first intend to implement an interface for advanced searches. Moreover, since our bitexts are already POS tagged, this information will be available soon. (6). The inclusion of POS tag information should allow much more complex and precise queries. In a later stage of development, we also intend to add word alignment. This alignment level is highly useful, since it allows the identification at a glance of the equivalent of a search term. Two word aligners are currently being tested: Giza ++ [26] and NATools.[27]

We are also very aware of some current shortcomings of our search engine, such as the non-availabiliy of KWIC format or the inflexibility of the results sorting. It should soon be possible to sort the segments according to different criteria, like alphabetical order, publication date, text type or equivalent of the search word or expression. For all these reasons, we are considering whether it will be possible to continue using a general search engine such as Solr or it would be preferable to switch to another more specific system, such as Corpus Workbench.[28]

Acknowledgement

The research project PaGeS Corpus has been funded by the Spanish Ministry of Economy and Competitiveness, Agencia Estatal de Investigación (FFI2013-42571-P, FFI2017-85938-R) and by the Galician Government (GI-1954 LitLinAl).

References

Čermák, Petr. This volume. InterCorp. Parallel corpus of 40 languages. In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* [Studies in Corpus Linguistics 90] Irene Doval & M. Teresa Sánchez (eds). Amsterdam: John Benjamins.

[26] <https://github.com/moses-smt/giza-pp>.

[27] <http://linguateca.di.uminho.pt/natools/>.

[28] <http://cwb.sourceforge.net/>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Čermák, František & Rosen, Alexandr. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 13(3): 411–427. https://doi.org/10.1075/ijcl.17.3.05cer

Clematide, Simon, Graën, Johannes & Volk, Martin. 2016. Multilingwis – A multilingual search tool for multi-word units in multiparallel corpora. In *Computerised and Corpusbased Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües*, Gloria Corpas Pastor (ed.), 447–455. Geneva: Tradulex.

Danielsson, Pernilla & Ridings, Daniel. 1997. Practical presentation of a Vanilla Aligner. In *TELRI Workshop in alignment and exploitation of texts, Ljubljana, Slovenia*. <http://www.kfben.com/dfilea/3122035922vanilla/ljubljana.pdf> (30 May 2017).

Dörk, Marian & Knight, Dawn. 2015. WordWanderer: A navigational approach to text visualisation. *Corpora* 10(1): 83–94. https://doi.org/10.3366/cor.2015.0067

Doval, Irene. 2016. PaGeS: Design and compilation of a bilingual parallel corpus German Spanish. *Epic Series in Languages and Linguistics* 1: 88–96. https://doi.org/10.29007/bcqd

Doval, Irene. 2017. POS-tagging a bilingual parallel corpus: Methods and challenges. *Research in Corpus Linguistics* 5: 35–46.

Doval, Irene. 2018. Das PaGeS-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache. *Revista de Filología Alemana* 26: 181–197. https://doi.org/10.5209/RFAL.60148

Łaziński, Marek & Kuratczyk, Magdalena. 2016 Korpus Polsko-Rosyjski Uniwersytetu Warszawskiego / The University of Warsaw Polish-Russian Parallel Corpus. In *Polskojęzyczne korpusy równoległe – Polish-language Parallel Corpora*, Ewa Gruszczyńska & Anieszka Leńko-Szymańska (eds), 83–95. Warszawa: Instytut Lingwistyki Stosowanej WLS, Uniwersytet Warszawski.

Lübke, Barbara & Liste Lamas, Elsa. 2019. *Raumrelationen im Deutschen: Kontrast, Erwerb und Übersetzung*. Tübingen: Stauffenburg.

Lüdeling, Anke & Kytö, Merja (eds). 2008. *Corpus Linguistics. An International Handbook*, Vol. 1. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110211429

Macken, Lieve, Trushkina, Julia, Paulussen, Hans, Rura, Lidia, Desmet, Piet & Wandeweghe, Wily. 2007. Dutch Parallel Corpus: A multilingual annotated corpus. In *Proceedings of the fourth Corpus Linguistics conference, University of Birmingham*. <http://ucrel.lancs.ac.uk/publications/CL2007/paper/173_Paper.pdf> (12 April 2017).

Molés-Cases, Teresa & Oster, Ulrike. This volume. Indexation and analysis of a parallel corpus using CQPweb: The COVALT PAR_ES corpus (EN/FR/DE>ES). In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* [Studies in Corpus Linguistics 90], Irene Doval & M. Teresa Sánchez (eds). Amsterdam: John Benjamins.

Rosen, Alexandr. 2016. InterCorp – a look behind the façade of a parallel corpus. In *Polskojęzyczne korpusy równoległe – Polish-language Parallel Corpora*, Ewa Gruszczyńska & Anieszka Leńko-Szymańska (eds), 21–40. Warszawa: Instytut Lingwistyki Stosowanej WLS, Uniwersytet Warszawski.

Steinberger, Ralf et al. 2006. The JRCAcquis: A multilingual aligned parallel corpus with 20 + languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. <https://arxiv.org/ftp/cs/papers/0609/0609058.pdf> (12 October 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Steinberger, Ralf et al. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation* 48(4): 679–707. https://doi.org/10.1007/s10579-014-9277-0

Tiedemann, Jörg. 2009. News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, Vol. V [Current Issues in Linguistic Theory 309], Nicolas Nicolov, Galia Angelova & Ruslan Mitkov (eds), 237–248. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.309.19tie

Tiedemann, Jörg. 2011. *Bitext Alignment*. San Rafael, CA: Morgan & Claypool Publishers.

Tóth, Krisztina, Farkas, Richárd & Kocsor, András. 2008. Sentence alignment of Hungarian–English parallel corpora using a hybrid algorithm. *Acta Cybern* 18: 463–478.

Varga, Dániel, Németh, László, Halácsy, Péter, Kornai, András, Trón, Viktor & Nagy, Viktor. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, 590–596.

Varga, Dániel. 2012. *Natural Language Processing of Large Parallel Corpora*. PhD dissertation. Budapest: Eötvös Loránd University.

Volk, Martin, Graen, Johannes & Callegaro, Elena. 2014. Innovations in parallel corpus search tools. In *Proceedings of LREC, Reykjavik*. <http://www.zora.uzh.ch/id/eprint/97282/1/Volk_Graen_Callegaro_LREC_2014_v06.pdf> (13 May 2017)

Volk, Martin, Clematide, Simon, Graen, Johannes, Ströbel, Phillip. 2016. Bi-particle adverbs, pos-tagging and the recognition of German separable prefix verbs. In *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), 296–305.

Wynne, Martin. 2008. Searching and concordancing. In *Corpus linguistics. An International Handbook*, Anke Lüdeling & Merja Kytö (eds), 706–737. Berlin: de Gruyter.

Zanettin, Federico. 2012. *Translation-driven Corpora*. London: Routledge.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Building EPTIC

A many-sided, multi-purpose corpus of EU parliament proceedings

Adriano Ferraresi and Silvia Bernardini
University of Bologna

This chapter describes the steps involved in the construction of EPTIC, an intermodal corpus of European Parliament speeches. Despite its limited size, this corpus has features that justify its labour-intensive building process, in particular its multiple alignments. The text-to-text alignments allow users to compare interpretations and translations of source speeches and their written-up reports, while text-to-video alignments allow them to access the multimedia components from concordance lines. To illustrate the potential of EPTIC, a case study is presented of English loan words in original, translated and interpreted Italian and French. Results suggest that borrowing is more likely to occur in translated Italian than in any of the other corpus components.

Keywords: intermodal corpora, text-to-text alignment, text-to-video alignment, corpus annotation, loan words

1. Introduction: Why another corpus of European Parliament speeches?

This chapter describes the steps involved in building a corpus of European Parliament (EP) speeches. One may wonder if this task is worth the effort, given that the language samples included in the corpus are among those most readily available to the research community, and generally viewed as inexpensive to acquire and process – as well as rather boring to analyse. Yet the similarities between EPTIC, the *European Parliament Translation and Interpreting Corpus*, and better known and widely used parallel corpus resources such as the *Europarl* corpus (Koehn 2005) are limited to their data source. As we shall see, EPTIC is a miniature corpus, but has features that make it unique, and that, we will argue, justify the time and effort spent building it.

https://doi.org/10.1075/scl.90.08fer
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

After describing the structure of the corpus at its current stage of development, and providing some basic data about its size and availability in Section 2, in Section 3 we detail the steps and tools used for its construction, and then briefly exemplify its potential for linguistic research in Section 4. We conclude with some forward-looking suggestions about the future of EPTIC and similarly complex, high-quality corpora.

2. What EPTIC looks like

2.1 One corpus, fourteen subcorpora

This section describes the basic structure of EPTIC with a focus on the subcorpora it includes and consequently the types of comparisons it affords. We shall consider one type of comparison at a time, relying on Zanettin's (2012: 11) typology of translation-driven corpora and progressing from simple and well-understood designs to more complex and innovative ones.

At its core, EPTIC can be described as a multilingual *parallel* corpus, since it includes text samples in one language and their translations into at least one other language. Secondly, at a higher level of complexity, EPTIC can also be seen as a multilingual *reciprocal* corpus, since more than one source language/target language pair is present, and the same language is both represented in the corpus as the *source* for translation and as its *target*. As is the case with other reciprocal corpora, such as COMPARA (Frankenberg-Garcia & Santos 2003) and the English Norwegian Parallel Corpus (Johansson 1998), EPTIC allows parallel comparisons (in two directions), monolingual comparable comparisons (of same-language texts that are either sources or targets of translation), and multilingual comparable comparisons (of texts in two languages sharing the same topic and genre).

Moving on to less mainstream corpus designs, EPTIC is also an *intermodal* corpus. This corpus type, proposed by Shlesinger (e.g. 2009) as an extension of Baker's (1995) original typology, is intended for the comparison of parallel or comparable samples of translated language and (transcribed) interpreted language. Each speech in EPTIC is indeed available in four different "versions": the transcription of the original delivery, the transcription of its interpretation, the official written-up version and its official translation, both published online as so-called "verbatim reports". Thus, we consider EPTIC a *quasi-parallel* corpus at the intermodal level, since the sources of the interpretations and those of the translations are not necessarily identical: they may differ slightly due to the editing done when producing the written reports. Given the presence of transcribed data and the inevitable loss of information that goes with it, multimedia support for the oral subcorpora is also provided, allowing users to access the video files directly from the concordance lines. This makes EPTIC *multimodal,* as well.

Lastly, to complicate things further, the multilingual, reciprocal, intermodal corpus design just described is currently applied to two language pairs in EPTIC, namely English<>Italian and English<>French, with the English sources being common to both subcorpora. All language pairs include English because, according to information gathered from the Directorate General for Translation, when producing the verbatim reports, English translations were carried out first, and used by all other translation departments as source texts. Therefore, unless one wishes to include cases of indirect translation, reciprocal *intermodal* corpora of EP speeches necessarily feature English.

Summing up, EPTIC has 14 separate subcorpora in 3 languages (English, French, Italian), 2 communication modes (written/spoken), and 2 mediation modalities (translated/interpreted), with access to time-synced videos. The subcorpora can be combined in different ways to provide multiple research perspectives on a single communicative event. The substantial effort needed to construct this corpus (more details in Section 3) is thus justified by the unique communicative situation it represents, one in which an authentic language event is subjected to both translation and interpreting between two dozen different languages: a dream data set not only for translation and interpreting scholars, but also for contrastive linguists and corpus linguists in general.

2.2 Practical details: Size and availability

Altogether the 14 subcorpora of EPTIC feature around 280,000 words. Table 1 shows size information about the corpus. Subcorpus names distinguish between source and target texts (*st*, *tt*), spoken components and written components (*sp, wr*), language of the corpus (*en* for English, *fr* for French and *it* for Italian), and, only for English target texts, their source language (*from-fr* and *from-it*).

As can be noticed, the English component of EPTIC is 1.3–1.5 times larger in terms of word tokens than the French and Italian ones. This is because we collect English target texts for both French and Italian source texts (see 2.1).

EPTIC is made freely available to the research community through the NoSketch Engine (henceforth NoSkE, Rychlý 2007).[1] This is a free web-based corpus query tool that supports sophisticated corpus queries based on contextual and linguistic annotation. What makes the NoSkE especially suitable as a query tool for EPTIC and similarly complex parallel corpora is its support for multiple

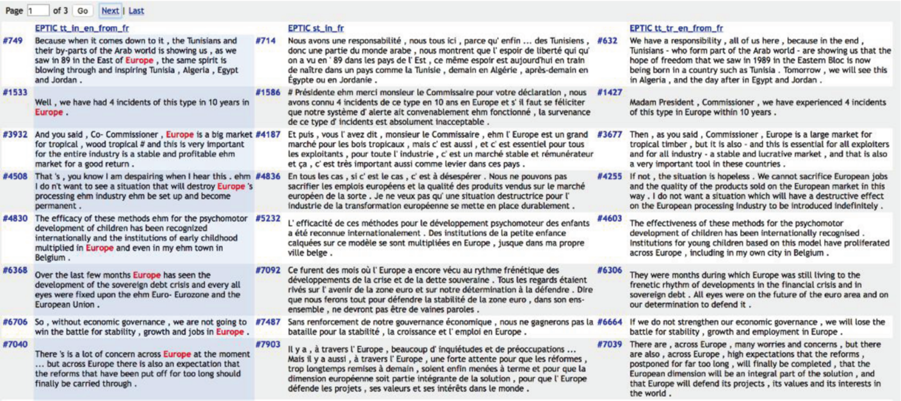
[1] The corpus is available via the "corpora" page of the CoLiTec research centre: <https://corpora.dipintra.it/eptic/>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 1. Size information on the EPTIC subcorpora

Language	Subcorpus	N. of texts	Word count
English	en-sp-st	64	21,106
	en-wr-st	64	20,091
	en-sp-tt_from-fr	65	20,836
	en-wr-tt_from-fr	65	21,641
	en-sp-tt_from-it	68	16,122
	en-wr-tt_from-it	68	17,775
	Sub-total	394	117,571
French	fr-sp-st	65	23,943
	fr-wr-st	65	23,159
	fr-sp-tt	63	20,558
	fr-wr-tt	64	22,940
	Sub-total	257	90,600
Italian	it-sp-st	68	16,586
	it-wr-st	68	16,521
	it-sp-tt	64	16,977
	it-wr-tt	64	19,665
	Sub-total	264	69,749
	TOTAL	**915**	**277,920**

parallel concordancing: given one query in a subcorpus, users can choose to visualize aligned sentences in all other subcorpora for which alignments are available (Figure 1; see also Section 3.3). What is more, concordance lines of transcriptions



	EPTIC tt_in_en_from_fr		EPTIC st_in_fr		EPTIC tt_tr_en_from_fr
#749	Because when it comes down to it , the Tunisians and their by-parts of the Arab world is showing us , as we saw in 89 in the East of Europe , the same spirit is blowing through and inspiring Tunisia , Algeria , Egypt and Jordan .	#714	Nous avons une responsabilité , nous tous ici , parce qu' enfin … des Tunisiens , donc une partie du monde arabe , nous montrent que l' espoir de liberté qui qu' on a vu en ' 89 dans les pays de l' Est , ce même espoir est aujourd'hui en train de naître dans un pays comme la Tunisie , demain en Algérie , après-demain en Égypte ou en Jordanie .	#632	We have a responsibility , all of us here , because in the end , Tunisians - who form part of the Arab world - are showing us that the hope of freedom that we saw in 1989 in the Eastern Bloc is now being born in a country such as Tunisia . Tomorrow , we will see this in Algeria , and the day after in Egypt and Jordan .
#1533	Well , we have had 4 incidents of this type in 10 years in Europe .	#1586	# Présidente ehm merci monsieur le Commissaire pour votre déclaration , nous avons connu 4 incidents de ce type en 10 ans en Europe et s' il faut se féliciter que notre système d' alerte ait convenablement ehm fonctionné , la survenance de ce type d' incidents est absolument inacceptable .	#1427	Madam President , Commissioner , we have experienced 4 incidents of this type in Europe within 10 years .
#3932	And you said , Co- Commissioner , Europe is a big market for tropical , wood tropical # and this is very important for the entire industry is a stable and profitable ehm market for a good return .	#4187	Et puis , vous l' avez dit , monsieur le Commissaire , ehm l' Europe est un grand marché pour les bois tropicaux , mais c' est aussi , et c' est essentiel pour tous les exploitants , pour toute l' industrie , c' est un marché stable et rémunérateur et ça , c' est très important aussi comme levier dans ces pays .	#3677	Then , as you said , Commissioner , Europe is a large market for tropical timber , but it is also - and this is essential for all exploiters and for all industry - a stable and lucrative market , and that is also a very important tool in these countries .
#4508	That 's , you know I am despairing when I hear this . ehm I do n't want to see a situation that will destroy Europe 's processing ehm industry ehm be set up and become permanent .	#4836	En tous les cas , si c' est le cas , c' est à désespérer . Nous ne pouvons pas sacrifier les emplois européens et la qualité des produits vendus sur le marché européen de la sorte . Je ne veux pas qu' une situation destructrice pour l' industrie de la transformation européenne se mette en place durablement .	#4255	If not , the situation is hopeless . We cannot sacrifice European jobs and the quality of the products sold on the European market in this way . I do not want a situation which will have a destructive effect on the European processing industry to be introduced indefinitely .
#4830	The efficacy of these methods ehm for the psychomotor development of children has been recognized internationally and the institutions of early childhood multiplied in Europe and even in my ehm town in Belgium .	#5232	L' efficacité de ces méthodes pour le développement psychomoteur des enfants a été reconnue internationalement . Des institutions de la petite enfance calquées sur ce modèle se sont multipliées en Europe , jusque dans ma propre ville belge .	#4603	The effectiveness of these methods for the psychomotor development of children has been internationally recognised . Institutions for young children based on this model have proliferated across Europe , including in my own city in Belgium .
#6368	Over the last few months Europe has seen the development of the sovereign debt crisis and every all eyes were fixed upon the ehm Euro- Eurozone and the European Union .	#7092	Ce furent des mois où l' Europe a encore vécu au rythme frénétique des développements de la crise et de la dette souveraine . Tous les regards étaient rivés sur l' avenir de la zone euro et sur notre détermination à la défendre . Dire que nous ferons tout pour défendre la stabilité de la zone euro , dans son ens- ensemble , ne devront pas être de vaines paroles .	#6306	They were months during which Europe was still living to the frenetic rhythm of developments in the financial crisis and in sovereign debt . All eyes were on the future of the euro area and on our determination to defend it .
#6706	So , without economic governance , we are not going to win the battle for stability , growth and jobs in Europe .	#7487	Sans renforcement de notre gouvernance économique , nous ne gagnerons pas la bataille pour la stabilité , la croissance et l' emploi en Europe .	#6664	If we do not strengthen our economic governance , we will lose the battle for stability , growth and employment in Europe .
#7040	There 's is a lot of concern across Europe at the moment … but across Europe there is also an expectation that the reforms that have been put off for too long should finally be carried through .	#7903	Il y a , à travers l' Europe , beaucoup d' inquiétudes et de préoccupations … Mais il y a aussi , à travers l' Europe , une forte attente pour que les réformes , trop longtemps remises à demain , soient enfin menées à terme et pour que la dimension européenne soit partie intégrante de la solution , pour que l' Europe défende les projets , ses valeurs et ses intérêts dans le monde .	#7039	There are , across Europe , many worries and concerns , but there are also , across Europe , high expectations that the reforms , postponed for far too long , will finally be completed , that the European dimension will be an integral part of the solution , and that Europe will defend its projects , its values and its interests in the world .

Figure 1. Multi-parallel EPTIC concordances in the NoSkE

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

can be linked to the corresponding bits of audios/videos of the original speeches (see Section 3.5).

3. Building EPTIC

3.1 Selecting and obtaining raw corpus materials

The European Parliament website[2] was the starting point of the EPTIC construction procedure, providing raw texts and videos, as well as contextual information on speeches and speakers. Through the dedicated interface, we searched for speeches in English, French and Italian delivered during plenary sessions in January and February 2011. To guarantee correct identification of the language in which speeches were originally delivered – a thorny issue for several EU Parliament-derived corpora (Nisioi et al. 2016) – we checked consistency of the information provided on the website against the videos themselves. The choice of dates was opportunistic: the practice of translating verbatim reports was unfortunately discontinued in the second half of 2011, therefore speeches were selected among the most recent for which translations are available (see Section 2.1).

For each speech thus identified, we collected three types of data:

- the verbatim reports in their original language, that is the written-up version of the speeches, which may or may not have undergone editing with respect to the original delivery;
- for the English reports, their translations into French and Italian, and for the French and Italian reports, their translations into English. Several EU officials consulted on the translation processes involving these documents reported that translations are entirely based on verbatim reports, with no reference to the original oral speech or the interpreters' renditions;
- the videos of the speeches,[3] which contain multi-audio tracks corresponding to the language in which the speech was delivered, and each of the languages into which it was rendered by the EP interpreters.

In addition to constituting the written components of EPTIC (translation source and target texts), the verbatim reports were used as a basis for transcribing its oral components, that is the speeches as delivered by speakers and interpreters. Details on the transcription procedures are provided in the next section.

[2] <http://www.europarl.europa.eu/plenary/en/debates-video.html>

[3] The videos can be downloaded from: <http://www.europarl.europa.eu/ep-live/en/plenary/>.

3.2 Transcribing the oral data

Starting from the verbatim reports, transcription of speeches and their interpretations was carried out based on the relevant audio track of the downloaded videos. Similar to the other interpreting corpora belonging to the EPIC suite (EPIC and EPICG, see Bernardini et al. 2018), the transcription is orthographic rather than phonetic (Niemants 2015), and follows the conventions of the EU *Interinstitutional Style Guide* concerning spelling, capitalization, acronyms and titles.[4]

Transcripts were segmented into sentence-like units by a full stop (or a question mark in the case of rising intonation), and include sub-sentential punctuation marks (commas), both inserted by the transcribers taking into account prosodic and syntactic cues. Although sentences are widely acknowledged not to be fully adequate for the segmentation of spoken language (Pietrandrea et al. 2014), especially in interpreting corpora, this information is essential to perform text-to-text and text-to-video alignment, as well as part-of-speech (POS) tagging (see 3.4). Applause, laughter and other conspicuous background noises were also inserted.

Finally, several markers of orality were incorporated in the transcriptions, including mispronunciations, truncated words/false starts and silent and filled pauses. In the case of mispronunciations, the normalized version of the word was included in the running text and the word as it was pronounced was included within slashes as an attribute of the first, so as not to inflate word counts. In case of unclear or inaudible stretches of text, a "#" was inserted. Table 2 summarizes the conventions just illustrated.

Table 2. Summary of transcription conventions adopted in EPTIC

Feature	Convention
Rising intonation	?
Background noises	[applause]
Sentence-like units	.
Sub-sentential units	,
Mispronunciations	Parlamento /Parlomento/
Truncated words	propo-
Silent pause	…
Filled pause	ehm
Inaudible segment	#

[4] <http://publications.europa.eu/code/en/en-000100.htm>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

3.3 Adding metadata

For each EPTIC text, a header was compiled containing information about the text itself, the context in which it originated and the EP speaker (and/or interpreter) who delivered it. Besides making it possible to relate "patterns of linguistic behavior […]

Table 3. Main types of metadata in EPTIC

Element	Attribute	Values and examples
TEXT	ID	e.g. 0001
	DATE	e.g. 17–01–11
	LENGTH	Short (<300 words) OR Medium (300–1000 words) OR Long (> 1000 words)
	LENGTH IN WORDS	[Minimum: 54, Maximum: 1556]
	DURATION	Short (<120 seconds) OR Medium (120–360 seconds) OR Long (>360 seconds)
	DURATION IN SECONDS	[Minimum: 24, Maximum: 597]
	SPEED	Slow (<130 w/m) OR Medium (131–160 w/m) OR High (>160 w/m)
	SPEED IN WORDS PER MINUTE (W/M)	[Minimum: 84.8, Maximum: 232.3]
	MODE OF DELIVERY	Read out OR impromptu OR mixed
	TOPIC	e.g. Politics
	SPECIFIC TOPIC	e.g. Order of business
SPEAKER	NAME	e.g. Tavares Rui
	GENDER	Male OR Female
	COUNTRY	e.g. Italy
	NATIVE	Yes OR No
	POLITICAL FUNCTION	e.g. MEP
	POLITICAL GROUP	e.g. GUE/NGL
SOURCE TEXT	LENGTH, LENGTH IN WORDS, DURATION, DURATION IN SECS., SPEED, SPEED IN W/M, MODE OF DELIVERY	*[Values duplicated from relevant source text]*
INTERPRETER	GENDER	Male OR Female
	NATIVE	Yes OR No

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

to [their] original habitat" (Burnard 2004: online), these metadata can be exploited to restrict queries based on attributes assigned to speakers and speech events.

The metadata encoded in the headers vary according to the type of texts considered. All headers contain information about the text (e.g. date, length, topic) and its originator (e.g. name, gender, whether native speaker of the language in which the speech was delivered). Oral texts also include data about their duration in seconds, speed and mode of delivery, and target texts include information about their sources (length, duration, speed and mode of delivery). Finally, interpreted texts include minimal information about the interpreter (gender and status as a native or non-native speaker of the target language). Table 3 illustrates the main types of metadata available in EPTIC.[5]

3.4 Performing text-to-text alignment

Alignment represents a crucial feature for a corpus like EPTIC, making it possible to fully exploit its potential as a multi-faceted source of data on translation and interpreting strategies. In what follows we describe the procedure employed to align EPTIC texts in a "traditional" parallel perspective (aligning source<>target texts), but also in an intermodal perspective (translations<>interpretations in the same language of the same source text), and a multilingual one (target texts for the same source text in languages A<>B). Alignment of transcripts to original audios/videos is described in Section 3.5.

One of the major problems we faced was the sheer number of alignments required. Each speech transcript had to be aligned to its interpretation transcript in one or two other languages (interpreting subcorpus), and each verbatim report to its translation in one or two other languages (translation subcorpus).[6] To account for the intermodal perspective, each transcript from the interpreting subcorpus had then to be aligned to its corresponding written version from the translation subcorpus. All in all, the resulting number of alignments exceeded two dozen.

While performing alignment with a fully automatic command-line tool like EasyAlign (Evert et al. 2016) was an enticing option, preliminary experiments revealed a less-than-ideal performance, especially in intermodal alignments, where one-to-one correspondences are rare.

The tool that we used instead was Intertext Editor (Vondřička 2014), an open source, user-friendly desktop aligner. Intertext Editor combines three welcome

[5] The full list of metadata is available on the "Documentation" page at: <https://corpora.dipintra.it/eptic/>.

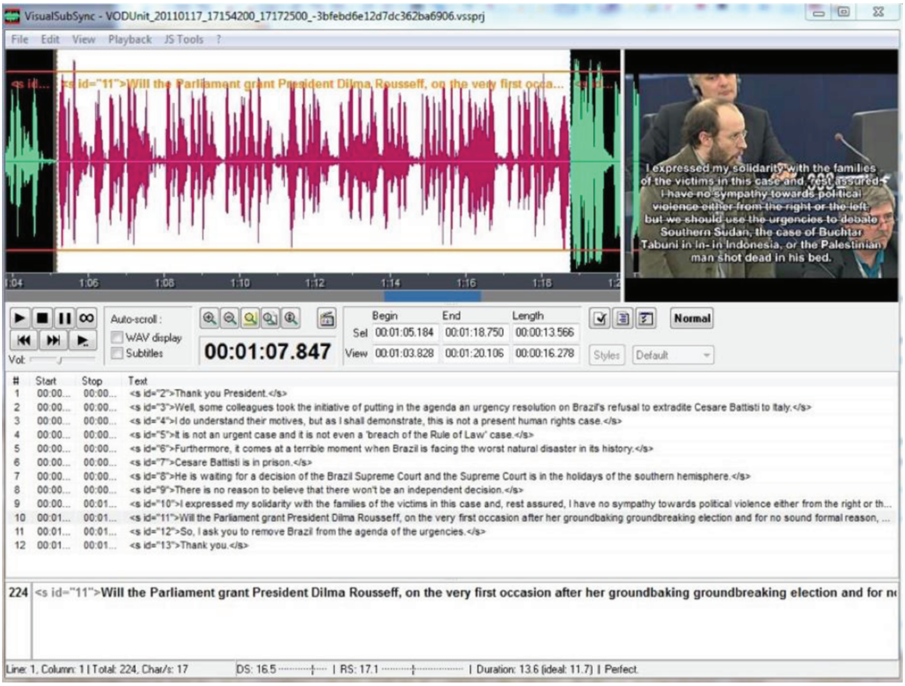
[6] English sources have French and Italian targets, while French and Italian sources only have English targets (see Section 2.1).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

features: first, it performs automatic pre-alignment of sentences using Hunalign (Varga et al. 2005), but also allows easy manual correction of misalignments. Second, it segments sentences based on user-defined regular expressions, and assigns a unique ID to each sentence, making it possible to specify multiple alignments across different subcorpora. Notice that for segmentation to be feasible, explicit sentence boundaries are necessary in the transcription of oral texts (see Section 3.2). Third, Intertext Editor provides several export options, including newline-aligned and TMX. Since multiple alignments are required, the default export format was used, which encodes alignment information as stand-off annotation. Three XML files are produced by Intertext Editor for each alignment: the segmented versions of both texts, and the alignment file showing the correspondences between the sentence(–like) units.

3.5 Performing text-to-video alignment

Text-to-video alignment is meant to provide access to the actual delivery of the speeches directly from concordance lines, making it possible to investigate prosodic or phonetic features that are necessarily lost in the transcription process.


VisualSubSync - VODUnit_20110117_17154200_17172500_-3bfebd6e12d7dc362ba6906.vssprj
File
Edit
View
Playback
JS Tools
?
<s id...
<s id="11">Will the Parliament grant President Dilma Rousseff, on the very first occa...
<s id...
1:04
1:06
1:08
1:10
1:12
1:14
1:16
1:18
1:2
Auto-scroll :
Vol:
Styles
Default
Normal

I expressed my solidarity with the families of the victims in this case and, rest assured, I have no sympathy towards political violence either from the right or the left but we should use the urgencies to debate Southern Sudan, the case of Buchtar Tabuni in In- in Indonesia, or the Palestinian man shot dead in his bed.

□ WAV display
□ Subtitles

00:01:07.847

	Begin	End	Length
Sel	00:01:05.184	00:01:18.750	00:00:13.566
View	00:01:03.828	00:01:20.106	00:00:16.278

```
#   Start     Stop      Text
1   00:00...  00:00...  <s id="2">Thank you President.</s>
2   00:00...  00:00...  <s id="3">Well, some colleagues took the initiative of putting in the agenda an urgency resolution on Brazil's refusal to extradite Cesare Battisti to Italy.</s>
3   00:00...  00:00...  <s id="4">I do understand their motives, but as I shall demonstrate, this is not a present human rights case.</s>
4   00:00...  00:00...  <s id="5">It is not an urgent case and it is not even a 'breach of the Rule of Law' case.</s>
5   00:00...  00:00...  <s id="6">Furthermore, it comes at a terrible moment when Brazil is facing the worst natural disaster in its history.</s>
6   00:00...  00:00...  <s id="7">Cesare Battisti is in prison.</s>
7   00:00...  00:00...  <s id="8">He is waiting for a decision of the Brazil Supreme Court and the Supreme Court is in the holidays of the southern hemisphere.</s>
8   00:00...  00:00...  <s id="9">There is no reason to believe that there won't be an independent decision.</s>
9   00:00...  00:01...  <s id="10">I expressed my solidarity with the families of the victims in this case and, rest assured, I have no sympathy towards political violence either from the right or th...
10  00:01...  00:01...  <s id="11">Will the Parliament grant President Dilma Rousseff, on the very first occasion after her groundbaking groundbreaking election and for no sound formal reason, ...
11  00:01...  00:01...  <s id="12">So, I ask you to remove Brazil from the agenda of the urgencies.</s>
12  00:01...  00:01...  <s id="13">Thank you.</s>
```

224 <s id="11">**Will the Parliament grant President Dilma Rousseff, on the very first occasion after her groundbaking groundbreaking election and for n**

Line: 1, Column: 1 | Total: 224, Char/s: 17   DS: 16.5 | RS: 17.1 | Duration: 13.6 (ideal: 11.7) | Perfect.

Figure 2. The VisualSubSync editor window. Each sentence (subtitle) is preceded by its ID

For this task, a minimalist approach was adopted in EPTIC that consisted in aligning the sentence-like segments also used for text-to-text alignment to their audio/video tracks using subtitling specialized translation software. Among the many freely available tools of this kind we selected VisualSubSync,[7] which in addition to providing standard functionalities to play a video and set start/end times for predefined segments, also displays the audio wave. This representation shows pauses in the form of a shallow line, thus facilitating the splitting of the audio track into sentence-like segments corresponding to their textual counterparts (see Figure 2). Start and end times were then easily converted into XML attribute values of the text segments themselves.

3.6 POS-tagging, lemmatization and indexing

The final phases of corpus construction consisted in annotating texts with POS tags and lemma information (using the TreeTagger),[8] integrating the different layers of annotation into a single file, and then indexing the corpus for consultation with the NoSkE (thanks to command line tools and ad hoc Perl scripts).

Figure 3 presents the final format of a text from the EPTIC English-from-Italian target oral (i.e. interpreted) corpus, ready for indexing. Specifically, it shows how the different layers of annotation are encoded in a mix of XML and vertical format. The text header, in XML, contains all the available contextual metadata. The body of the text, that is the actual transcript (or verbatim report for the written components of EPTIC) is instead set in the vertical format produced by the TreeTagger, except for information on sentences (the "s" tag), which is taken from the XML output of Intertext Editor.

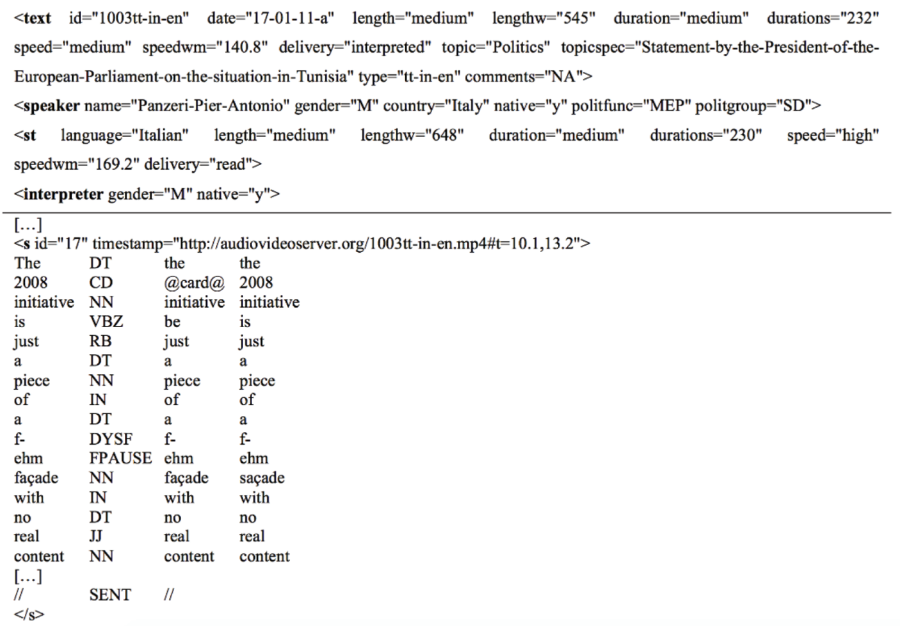
The first three columns in the text body encode, respectively, the normalized text, the POS of each word, and its lemma; the fourth column contains the non-normalized text (see Section 3.2). Concerning POS, three tags were added to the TreeTagger default tagset to account for the specificities of the spoken components of EPTIC: DYSF indicates dysfluencies, while FPAUSE and EPAUSE indicate filled and empty pauses respectively.

The "s" element is used both for text alignment and to encode information on text-to-video alignment. In particular, the "timestamp" attribute exploits a convenient feature of the NoSkE that makes it possible to link URLs to each sentence: this feature is used as a workaround to the lack of support for integration of audio and video files into the NoSkE interface. For each sentence in the EPTIC spoken

[7] <http://www.visualsubsync.org>.

[8] <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



```
<text id="1003tt-in-en" date="17-01-11-a" length="medium" lengthw="545" duration="medium" durations="232"
speed="medium" speedwm="140.8" delivery="interpreted" topic="Politics" topicspec="Statement-by-the-President-of-the-
European-Parliament-on-the-situation-in-Tunisia" type="tt-in-en" comments="NA">
<speaker name="Panzeri-Pier-Antonio" gender="M" country="Italy" native="y" politfunc="MEP" politgroup="SD">
<st language="Italian" length="medium" lengthw="648" duration="medium" durations="230" speed="high"
speedwm="169.2" delivery="read">
<interpreter gender="M" native="y">
[…]
<s id="17" timestamp="http://audiovideoserver.org/1003tt-in-en.mp4#t=10.1,13.2">
The        DT       the        the
2008       CD       @card@     2008
initiative NN       initiative initiative
is         VBZ      be         is
just       RB       just       just
a          DT       a          a
piece      NN       piece      piece
of         IN       of         of
a          DT       a          a
f-         DYSF     f-         f-
ehm        FPAUSE   ehm        ehm
façade     NN       façade     saçade
with       IN       with       with
no         DT       no         no
real       JJ       real       real
content    NN       content    content
[…]
//         SENT     //
</s>
```

Figure 3. EPTIC final format, ready for indexing with the NoSkE

subcorpora, a URL is provided pointing to the server hosting the videos and running an ad hoc PHP script to display them directly in the browser; the final part of the URL (e.g., "&start=00:10.1&end=00:13.2") specifies the start and end time of the sentence, as determined during the text-audio/video alignment phase. When clicking on this link from concordance lines, a new browser window opens and the matching fragment of the video/audio file is played within the browser itself.

4. An example: English loan words in Italian and French

Having detailed the process of building EPTIC, in this section we illustrate its potential through a simple example. As is well-known, English is the de facto lingua franca both in Europe in general and at the European Union institutions. Even though the right for each member of the EP to use their country official language is guaranteed, many, particularly commissioners, choose to speak English (Codrea-Rado 2014). This situation is in fact similar to other, less privileged transnational communities of practice across Europe (Motschenbacher 2013). One might therefore hypothesize that English loan words easily make their way into other languages, and indeed this feature defines that sort of EU-jargon that Italian linguists sometimes disparagingly call "*Europeese*", that is a variety of Italian filled with obscure, technical words often borrowed from English.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Using EPTIC, one could compare the use of loan words in Italian and in French, a language known for its resistance to foreignisms (Bogaards 2008). Even though the corpus is very small, its various components are of similar sizes and composition, and therefore easily comparable. Focusing on the transcripts of speeches originally delivered in Italian, extracting all noun types (rather than tokens) that do not end in a vowel, and ignoring fully integrated borrowings such as *standard* and *sport*, we find 7 English expressions (see Table 4). The French speeches contain a similar number of loans (6), but one is a quotation (*much ado about nothing*), three are almost entirely integrated (*jobs*, *tests*, *budget*), and one is a *faux anglicisme,* a pseudo borrowing, according to the *Larousse* dictionary (*tennisman*).[9] The only other loan word, which is also present in the Italian corpus, is *lady,* a term used when addressing Baroness Catherine Ashton. No *Europeese*-like expressions (like *road map* or *stress test*) are found in the French source speech corpus, confirming the general perception of Italian as more permeable to borrowings than French.

Table 4. Borrowings from English in Italian and French source speeches

Italian	French
1. road map	1. budget
2. stress test	2. lady
3. standby	3. jobs
4. lobby	4. much ado about nothing
5. partner	5. tests
6. lady	6. tennisman
7. business	

We may wonder if this trend also applies to written communication, that is whether the use of anglicisms remains constant, increases or decreases in the verbatim reports (the written-up versions of the speeches). When the same searches are conducted on the Italian written source component of EPTIC, the same expressions are found. In the corresponding French corpus, instead, one difference emerges: the English loan word *jobs* is replaced by *emplois*. A single example does not tell us much, and yet, since the written verbatim reports tend to reproduce the uttered speeches as exactly as possible, this instance of gallicization does not seem accidental.

Taking the argument further, one might hypothesize that the real source of *Europeese* is not (only) the Italian of Italian members of the EP, but also the Italian

[9] <http://www.larousse.fr/dictionnaires/francais>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

of interpreters and translators. Since interference has been repeatedly suggested to be a typical feature (or universal) of translation broadly conceived (Toury 1995; Chesterman 2004), it is not unlikely that interpretations and translations contain an even higher number of anglicisms than speeches originally delivered in Italian. Table 5 lists the loan words found in Italian interpreted and translated from English, once again comparing results with those found in the comparable French components of EPTIC.

Table 5. Anglicisms in mediated Italian and French

Interpreted Italian	Translated Italian	Interpreted French	Translated French
partner	partner	budget	stocks
leader	leader	budgétaire	tests
follow-up	test	budgétaires	budget
leadership	camp	e-mails	budgets
leaders	establishment	leader	budgétaire
roadmap	e-mail	leaders	budgétaires
stakeholder	leadership	tests	week-end
stock	partnership		
team	roadmap		
	stock		
	team		
	workshop		

As can be seen, mediated language does contain a number of borrowings slightly higher than non-mediated (source) speeches/texts. Interestingly, translated Italian contains even more English borrowings than interpreted Italian, suggesting that their use is probably not due to the difficulty of blocking interference when processing language in real-time, as happens in simultaneous interpreting. The comparison with French shows no such difference between interpreting and translation: similar numbers of borrowings are found.

Finally, by extracting and analyzing parallel concordances, one could try to (dis)confirm the hypothesis that the use of loan words is indeed due to interference. While this may seem almost a corollary, in fact the very existence of *Europeese* could also result in a tendency for EU professionals to use English words in their peculiar variety of Italian, even when the same word is not there in the source speech/text. In Chesterman's words, the use of loan words could equally be due to an S-universal (interference) and a T-universal (normalization, or respect of target language norms). The only way to find out is to look at the parallel concordances.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

For the sake of brevity, only parallel concordances taken from the Italian/English translation corpus are shown in Table 6. When more than one occurrence is found, only the first is shown.

In most cases, English words in Italian translations are transferred directly from the sources, but in 5, a French loanword in English (*elite*) is translated into Italian with an English loan word (*establishment*). This might be taken to suggest that translators make a conscious choice when they let the English words through. Together with the higher presence of borrowings in translation than in interpretation, this observation could be taken to point more towards normalization than interference.

Of course, this is only an example (not even a case study), and consequently no firm conclusions can be drawn. But we hope to have illustrated some of the potential of EPTIC, which resides mainly in its multi-faceted yet compact structure, in which the different components illuminate each other.

Table 6. Anglicisms in Italian translations and their sources

1a. Unione europea e dai suoi *partner* 1b. the European Union and its *partners*
2a. i *leader* politici riconoscessero 2b. political *leaders* woke up to the realities
3a. Alcuni *test* effettuati dal ministero 3b. Some *test* results from the Agriculture Ministry
4a. al campo di al-Hol *camp* in Siria 4b. go to the al-Hol *camp* in Syria
5a. I compiti dell'*establishment* e della società 5b. The tasks faced by the *elite* and by society
6a. ricevendo numerose *e-mail* da parte di cittadini 6b. getting lots of *e-mails* from concerned citizens
7a. La nuova *leadership* politica tunisina deve 7b. Tunisia's new political *leadership* also needs
8a. relativi al lavoro di *partnership* 8b. about *partnership* working
9a. alla *roadmap* ed alla questione 9b. the *road map* and the issue
10a. gli effetti sugli *stock* ittici globali 10b. the effects on global fish *stocks*
11a. un *team* dell'Ufficio alimentare 11b. a *team* from the Food and Veterinary Office
12a. l'organizzazione di un *workshop* 12b. a *workshop*

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

5. Conclusion: Teaming up

This contribution has described EPTIC, a tiny but highly complex trilingual, parallel, intermodal and multimodal corpus, focusing in particular on the procedure involved in building it. To illustrate some of its potential, a small-scale study on anglicisms in mediated (translated and interpreted) and non-mediated ("original") Italian and French was carried out. As the description and exemplification have hopefully suggested, EPTIC lends itself to many research purposes, such as the investigation of register variation across speech and writing, of contrastive differences between languages, of typical features of mediated vs. non-mediated language, of typical features of translated vs. interpreted language, in all the languages it features (currently English, French and Italian).

The greatest limits at the moment are the small size of the corpus and the small number of languages represented (3 out of the 23 available ones). To increase the size of EPTIC and the number of languages covered, a community effort is needed, as no single team is likely to have the expertise and resources required. English<>Finnish, English<>Slovene and English<> Polish subcorpora are currently being constructed, and we hope that other interested linguists from diverse language backgrounds join us in the development of EPTIC. As claimed by Granger (2010: online), "[w]e need more and better corpora for crosslinguistic research and as data collection is very time-consuming, there is a great deal to be gained from joining forces".

Acknowledgement

The work described in this contribution is the result of collaboration with Maja Miličević Petrović (University of Belgrade) and Marie-Aude Lefer (Université catholique de Louvain).

References

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2): 223–243. https://doi.org/10.1075/target.7.2.03bak

Bernardini, Silvia, Collard, Camille, Ferraresi, Adriano, Russo Mariachiara & Defrancq, Bart. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In *Making Way in Corpus-based Interpreting Studies*, Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds), 21–42. Singapore: Springer. https://doi.org/10.1007/978-981-10-6199-8_2

Bogaards, Paul. 2008. *On ne parle pas franglais: La langue française face à l'anglais*. Brussels: De Boeck/Duculot.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Burnard, Lou. 2004. Metadata for corpus work. In *Developing Linguistic Corpora: A Guide to Good Practice*, Martin Wynne (ed.). <http://ota.ox.ac.uk/documents/creating/dlc/> (30 June 2017).

Chesterman, Andrew. 2004. Hypotheses about translation universals. In *Claims, Changes and Challenges in Translation Studies* [Benjamins Translation Library 50], Gyde Hansen, Kirsten Malmkjaer & Daniel Gile (eds), 1–13. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.50.02che

Codrea-Rado, Anna. 2014. European parliament has 24 official languages, but MEPs prefer English. *The Guardian*. <https://www.theguardian.com/education/datablog/2014/may/21/european-parliament-english-language-official-debates-data> (30 October 2017).

Evert, Stefan & the CWB Development Team. 2016. The IMS Open Corpus Workbench (CWB) Corpus Encoding Tutorial. CWB Version 3.4: <http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial/> (30 October 2017).

Frankenberg-Garcia, Ana & Santos, Diana. 2003. Introducing COMPARA: The Portuguese–English parallel corpus. In *Corpora in Translator Educatio*, Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds), 71–87. Manchester: St. Jerome.

Granger, Sylviane. 2010. Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Journal of Shanghai Jiaotong University* 2: 14–21.

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

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, 79–86. Phuket, Thailand.

Motschenbacher, Heiko. 2013. *New Perspectives on English as a European Lingua Franca*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.182

Niemants, Natacha. 2015. Transcription. In The Routledge Encylopedia of Intepreting Studies, Franz Pöchhacker (ed), 421–422. London: Routledge.

Nisioi, Sergiu, Rabinovich, Ella, Dinu, Liviu P. & Wintner, Shuly. 2016. A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4197–4201.

Pietrandrea, Paola, Kahane, Sylvain, Lacheret-Dujour, Anne & Sabio, Frédéric. 2014. The notion of sentence and other discourse units in corpus annotation. In *Spoken Corpora and Linguistic Studies* [Studies in Corpus Linuistics 61], Tommaso Raso & Heliana Mello (eds), 331–364. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.61.12pie

Rychlý, Pavel. 2007. Manatee/Bonito – A modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70. Masaryk University, Brno.

Shlesinger, Miriam. 2009. Towards a definition of interpretese: An intermodal, corpus-based study. In *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile* [Benjamins Translation Library 80], Gyde Hansen, Andrew Chesterman & Heidrun Gerzymisch-Arbogast (eds), 237–253. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.80.18shl

Toury, Gideon. 1995. *Descriptive Translation Studies – and Beyond* [Benjamins Translation Library 4]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.4

Varga, Dániel, Németh, László, Halácsy, Péter, Kornai, András, Viktor Trón & Nagy, Viktor. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP* 2005, 590–596.

Vondřička, Pavel. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 1875–1879.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Zanettin, Federico. 2012. *Translation-driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Abingdon: Taylor & Francis.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Enriching parallel corpora with multimedia and lexical semantics

From the CLUVI Corpus to WordNet and SemCor

Xavier Gómez Guinovart
University of Vigo

In this chapter, I present the main characteristics of the CLUVI Corpus, an open collection of sentence-level aligned parallel corpora with over 44 million words in nine specialised domains (fiction, computing, popular science, biblical texts, law, consumer information, economy, tourism, and film subtitling) and different language combinations including Galician, Spanish, English, French, Portuguese, Catalan, Italian, Basque and Latin. Then, I present the methodology developed for extending the film subtitles section of the CLUVI Corpus with multimedia data. Finally, I discuss the resources and methods used to build the SensoGal Corpus, a SemCor-based English-Galician parallel corpus semantically annotated based on WordNet and aligned at the sentence and word levels.

Keywords: parallel corpora, multimedia, lexical semantics, WordNet, SemCor

1. Introduction

Parallel corpora are digitized collections of texts stored in both their original and translated version. Although it is possible to incorporate any type of linguistic information into parallel corpora, in general, parallel corpora are usually annotated with information on the equivalences of translation between the segments (words, phrases, sentences or other textual units). This process is called alignment, and these enriched parallel corpora are called aligned parallel corpora (Véronis 2000). Aligned parallel corpora have numerous applications in multilingual natural language processing, in the fields of machine translation (Koehn 2005), translation memories (Keshtkar & Mosavi Miangah 2012), lexical and terminological extraction (Tufiş 2007), second language teaching (Montero Perez et al. 2014), and contrastive linguistics (Santos 2004), among others. For a recent review of the state

https://doi.org/10.1075/scl.90.09gom
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

of the art of this field, including an up-to-date survey of available parallel corpora, see the work of Mikhailov & Cooper (2016).

This chapter provides an overview of the research on parallel corpora carried out for more than a decade by the Seminario de Lingüística Informática (SLI) at the Universidade de Vigo in Galicia, Spain, with a focus on the development and exploitation of the CLUVI Corpus and on its multimedia extensions. I will also consider the most recent research on how to incorporate the advances into the design and study of computational lexical semantics to parallel corpora.[1]

2. The CLUVI Corpus

2.1 Corpus description

The CLUVI Corpus[2] is an open collection of human-annotated sentence-level aligned parallel corpora originally designed to cover specific areas of the contemporary Galician language in relation to other languages. With over 44 million words, the CLUVI collection currently comprises twenty parallel corpora in nine specialized registers or domains (fiction, computing, popular science, biblical texts, law, consumer information, economy, tourism, and film subtitling) and different language combinations with Galician, Spanish, English, French, Portuguese, Catalan, Italian, Basque, German and Latin. The coverage and size of the CLUVI Corpus are shown in Table 1, where the different subcorpora in the CLUVI collection are listed along with data about their current status.

At this moment, the CLUVI is the parallel corpus that contains the greatest number of translations from/to the Galician language and the widest thematic collection. Galician texts in CLUVI sum up about 11,000,000 words, which means a quarter of the total of the tokens in the corpus for the 10 languages gathered, and are representative of 9 types of specialized translation: legal, biblical, scientific-technical, literary, software localization, consumer information, film subtitling, economy and tourism.

[1] This research has been carried out in the framework of the TUNER project (TIN2015–65308-C5–1-R) supported by the Ministry of Economy and Competitiveness of the Spanish Government and the European Fund for Regional Development (MINECO/FEDER).

[2] International Standard Language Resource Number (ISLRN): 910–993–402-072-9.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 1. Coverage and size of the CLUVI Corpus

Language combinations and domains	Words
LEGA: Corpus of Galician–Spanish legal texts	6,582,415
BIBLOGAL: Corpus of Latin–Galician–Brazilian Portuguese–European Port–Catalan–Italian–Spanish–English biblical texts	5,489,607
UNESCO: Corpus of English–Galician–French–Spanish scientific-technical divulgation	3,724,620
LOGALIZA: Corpus of English–Galician software localization	3,706,242
CONSUMER: Corpus of Spanish–Galician–Catalan–Basque consumer information	5,586,431
TECTRA: Corpus of English–Galician literary texts	2,465,154
FEGA: Corpus of French–Galician literary texts	1,863,959
DEGA: Corpus of German–Galician literary texts	366,038
GALEA: Corpus of Galician–Spanish literary texts	162,795
PEGA: Corpus of Portuguese–Galician literary texts	68,431
EGAL: Corpus of Galician–Spanish economy texts	718,642
TURIGAL: Corpus of Spanish–Galician tourism texts	325,389
VEIGA: Corpus of English–Galician film subtitling	294,714
LEGE-BI: Corpus of Basque–Spanish legal texts	2,384,053
LOGALIZA: Corpus of English–Spanish software localization	4,992,133
TECTRA: Corpus of English–Spanish literary texts	2,108,141
PALOP: Corpus of Portuguese–Spanish postcolonial literature	566,590
TURIGAL: Corpus of Portuguese–English tourism texts	1,285,764
TECTRA: Corpus of English–Portuguese literary texts	875,595
SCOPE: Corpus of English–Spanish economy texts	1,151,544
CLUVI Corpus total size:	**44,163,783**

Other parallel corpus that currently facilitate access to translations to Galician are the Opus Corpus[3] (Tiedemann 2012) and the Per-Fide Corpus[4] (Almeida et al. 2014). On the one hand, the Opus collection provides Galician translated texts, mainly from English, taken from the web and automatically aligned at sentence level; it has an extension of around 7,600,000 tokens in Galician, and is taken primarily from the localization of the Linux operating system environment (Gnome, KDE4 and Ubuntu). On the other hand, the Per-Fide collection includes

[3] Available at <http://opus.lingfil.uu.se/>.

[4] Available at <http://per-fide.di.uminho.pt/>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Portuguese–Galician software localization parallel texts derived from the Opus with about 400,000 tokens in Galician.

The format chosen for storing the aligned parallel texts in CLUVI is an adaptation of the TMX format (Savourel 2005), as this is the XML encoding standard for translation memories, regardless of the application used. A translation memory is a database that collects and records source text segments and their corresponding translated versions with the purpose of being reused for further translations via a computer-aided translation system. Albeit with some differences, an aligned parallel corpus is equivalent to a translation memory. Indeed, the last few years have seen an increasing number of TMX-encoded aligned parallel corpora, which offer the additional advantage that they can be used as translation memories for feeding computer-aided translation programs (Simões/Gómez Guinovart & Almeida 2004).

Since 2003, the SLI offers the possibility of searching and browsing the CLUVI parallel corpora online.[5] The parallel corpora managed by the web application are stored in the XML CLUVI specification, whereas the searching and browsing tool designed in PHP was specifically created to carry out bilingual searches in tagged texts that are conformant to this specification (Gómez Guinovart & Sacau Fontenla 2004b). This search application allows for very complex searches of isolated words or sequences of words, and shows the bilingual equivalences of the terms in context, as they appear in real and referenced translations. Due to copyright issues, it returns a maximum of 1,500 hits only, in order not to exceed the limits of the right to quote. Users can search for terms in either language of the corpus, although it is also possible to carry out true bilingual searches, that is, to search for bilingual segments that simultaneously contain a term in the source language and another term in the target language. Search results are displayed in a parallel fashion as a list of translation units. In addition, the LEGA Corpus from CLUVI can be downloaded via MetaShare[6] with CC BY-NC-SA 3.0 license.

2.2 Tagging the CLUVI Corpus

The basic segmentation unit for the alignment of the CLUVI parallel texts is the orthographic sentence of the source text. Therefore, the correspondence between source and target text will always be of the 1:n type. Frequently one sentence of the source text corresponds with one sentence of the translation (1: 1). Nevertheless, there are cases in which a source sentence is not translated (1: 0), or in which a

[5] The CLUVI online is available at <http://sli.uvigo.gal/CLUVI/>.

[6] Available at <http://hdl.handle.net/10230/20051>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

source sentence corresponds with half a sentence (1: 1/2) or with two or more sentences of the translation (1: 2, 1: 3…), or even in which a sentence of the translation does not correspond with any source sentence (0: 1). Moreover, translating sometimes implies movements of sentences, or movements of source fragments from their original sentences to other sentences in translation. These movements are reordered in the target section of CLUVI parallel corpora to fit with the 1:n alignment criterion that preserves the integrity and the order of the translation units of the source text. This criterion is crucial when applied to the processing of multilingual corpora, where source sentences must permit to establish correspondences among equivalent sentences in various languages.

The TMX specification does not consider the encoding of these stylistic aspects of translations, because it has been designed for the storage and exchange of translation memories, and not for the representation of equivalent segments in parallel corpora. The TMX-based CLUVI encoding system uses an adapted version of some tags which are part of the TMX 1.4 specification (Savourel 2005) in order to represent the non-1: 1 correspondences and reorderings encoded in the CLUVI parallel corpora. All these stylistic aspects of the corpus can be annotated according to the TMX-based CLUVI Corpus XML specification for parallel corpora which is summarized in Figure 1.

```
<!-- CLUVI_TMX DTD -->
<!ELEMENT cluvi_tmx (header, body) >
<!ATTLIST cluvi_tmx
      version CDATA #REQUIRED >
<!ELEMENT header (#PCDATA)>
<!ELEMENT body (tu*) >
<!ELEMENT tu (tuv+) >
<!ELEMENT tuv (seg) >
<!ATTLIST tuv
      xml:lang CDATA #REQUIRED>
<!ELEMENT seg (#PCDATA | hi | ph)*>
<!ELEMENT hi (#PCDATA)>
<!ATTLIST hi
      type CDATA #IMPLIED
      x CDATA #IMPLIED>
<!ELEMENT ph EMPTY>
<!ATTLIST ph
      x CDATA #IMPLIED>
```

Figure 1. TMX-based CLUVI Corpus XML specification

The stylistic aspects of translation encoded in the CLUVI corpora can be described as either omissions, additions or reorderings, and will be tagged using an adapted version of the TMX 1.4 content elements <*hi* > (or highlight) and < *ph* > (or placeholder). An omission occurs when an item of the source text does not correspond

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

with any item of the target text, that is, when a sentence or part of a sentence is not translated. Omissions in the CLUVI parallel corpora are encoded by means of the <*hi* > element. According to the TMX 1.4 specification, the <*hi* > element “delimits a section of text that has special meaning, such as a terminological unit, a proper name, an item that should not be modified, etc. It can be used for various processing tasks” (Savourel 2005). In the TMX-based CLUVI encoding, the <*hi* > element marks the piece in the source text that is omitted in the target text. This use of the <*hi* > tag is noted by means of the type attribute with the *“supr”* (deleted) value. For instance, the English–Galician aligned sentences in (1) would be encoded as the translation unit in (2).

(1) ‘Hello,’ I said. [English] / -Ola. [Galician]
< tu> < tuv xml:lang = “en” > <seg>‘Hello,’<hi type = “supr” > I said.</
hi > </seg> </tuv> < tuv xml:lang = “gl” > <seg>-Ola.</seg> </tuv> </tu>

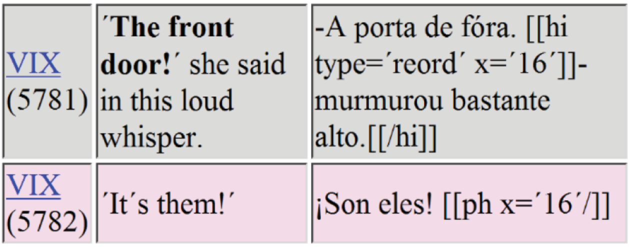
On the other hand, the translation technique known as addition involves the insertion of elements in the target text that have no correspondence in the source text. Addition is also encoded in the CLUVI by means of the <*hi* > element, which highlights the inserted unit in the target text. This use of the <*hi* > tag is indicated by means of the type attribute with the *“incl”* (included) value. The added text joins the translation unit into which it is inserted. If the new element is a sentence (or a sequence of sentences), it joins either the preceding or the following translation unit, depending on its context, thus respecting the 1:n alignment criterion. For instance, the alignment in (3) would be encoded as (4).

(2) ‘Hello.’ / -Ola – dixen.
< tu> < tuv xml:lang = “en” > <seg>‘Hello.’</seg> </tuv> < tuv
xml:lang = “gl” > <seg>-Ola < hi type = “incl” > – dixen.</hi > </tuv> </tu>

The reordering in translation implies movements of sentences, or movements of source fragments from their original sentences to other sentences in translation. These movements are reordered in the target text to fit with the 1:n alignment criterion that preserves the integrity and the order of the translation units of the source text. Reordering is encoded in CLUVI by means of a combination of the <*hi* > element and the <*ph* > element. The phrase or sentence moved is tagged with a < *hi* > element, with a type attribute with the *“reord”* value, as well as with an *x* attribute with a numeric value acting as an unambiguous index. Moreover, the place in the texts from where the segment was moved is indicated by means of a < *ph* > element. According to the TMX 1.4 specification, the <*ph* > element is used “to delimit a sequence of native standalone codes in the segment. Standalone codes are codes that are not opening or closing of a pair, for example empty elements in XML” (Savourel 2005). In the TMX-based CLUVI encoding, the adapted

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

<*ph* > element marks the departure point of the movement, and the relationship between the element moved and its place of origin is encoded in the <*ph* > element by means of an *x* attribute that shares its value with the index encoded in the <*hi* > element of the segment moved. The example in (5–6) shows how reorderings are encoded, and Figure 2 illustrates how the alignments are displayed by the web application when actually searching the corpus.



VIX (5781)	′**The front door!**′ she said in this loud whisper.	-A porta de fóra. [[hi type=′reord′ x=′16′]]-murmurou bastante alto.[[/hi]]
VIX (5782)	′It′s them!′	¡Son eles! [[ph x=′16′/]]

Figure 2. Reordering in CLUVI

(3) ‘The front door!’ she said in this loud whisper. ‘It’s them!’ / -A porta de fóra. ¡Son eles! – murmurou bastante alto.

(4) <tu> < tuv xml:lang = “en” > <seg>‘The front door!’ she said in this loud whisper.</seg> </tuv> < tuv xml:lang = “gl” > <seg> – A porta de fóra. < hi type = “reord” x = “16” > – murmurou bastante alto.</hi > </seg> </tuv> </tu> <tu> < tuv xml:lang = “en” > <seg>It’s them.</seg> </tuv> < tuv xml:lang = “gl” > <seg>¡Son eles! < ph x = “16”/></seg> </tuv> </tu>

2.3 Extending the CLUVI Corpus with multimedia data

Multimedia parallel corpora are a relatively scarce resource in the field of language technologies, due to the problem of obtaining translated multimedia materials, the difficulties of transcription and the technical complications of their processing. In turn, their use allows us to analyze aspects of multimedia translation that would be impossible to study from a merely textual perspective. In this section, I will present the methodology developed by the SLI for extending the CLUVI Corpus with multimedia data, focusing on the building of a multimedia extension of the VEIGA Corpus (Sotelo Dios & Gómez Guinovart 2012). The VEIGA Corpus is an English–Galician corpus consisting of 24 American, British, and Australian films subtitled in both English (intralingual subtitling) and Galician (interlingual subtitling) for DVD, cinema and Internet distribution. Developed under the broader framework of the CLUVI Corpus, VEIGA was born as a text-only corpus of subtitles. It was not until recently that we decided to make it multimedia, as soon as we found the appropriate tools to process the data and to make it accessible to the

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

public in what we considered to be an appropriate way. The VEIGA multimedia corpus of subtitles is currently available for public consultation at the CLUVI site.[7] However, it should be noted that only 13 of the 24 films are available in multimedia format at the time of writing.

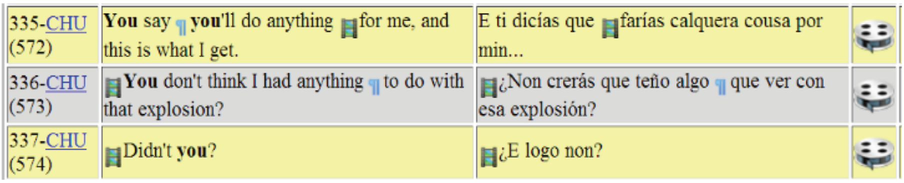
The CLUVI Corpus functions as a repository of parallel corpora of different sizes and thematic fields, all of which undergo identical compiling and processing routines, and can be similarly accessed from one single search interface. Nonetheless, the VEIGA Corpus requires further processing in comparison to the other CLUVI corpora. Besides annotating stylistic aspects of translation such as omissions, additions and reordering of translation units, all the subtitles include both the in-cue and out-cue time (the time markers indicating where each subtitle should appear and disappear) and the line break indicator, allowing users to examine aspects which are inherent to subtitling practice, for example time and space constraints, segmentation, and condensation, among other specificities. In addition to this, the multimedia version of VEIGA enables users to stream the video clips corresponding to the bilingual pairs found in the search results, thus giving them access to the (co-)text in its original, multi-semiotic form. This means that wherever there is a result that matches the query in text format, the search interface shows a link to the corresponding video clips subtitled in each of the two languages involved (English and Galician) as shown in Figures 3–5.

All the above mentioned aspects of the VEIGA Corpus are annotated according an extended CLUVI XML specification, which is summarized in Figure 6.

Tagging the VEIGA Corpus at the textual/audiovisual interface level implies, on the one hand, tagging the correspondences between the English subtitles stored as XML textual data in the extended CLUVI XML specification and the equivalent segment of the original English-language film with English subtitles, and, on the other hand, tagging the correspondences between the Galician subtitles stored as XML textual data and the equivalent segment of the original English-language film with Galician subtitles. In order to be able to establish these textual/audiovisual correspondences, all of the VEIGA English-language films have been cut into video clips, each one corresponding to a subtitle. A first step is to check if the subtitles are in sync with the movie. In some cases, mostly when the subtitle file and the movie come from different sources, we need to edit the subtitles and add a time delay (forward or backward) so that their speed matches that of the video. Secondly, we embed the subtitles in the two languages in the original film. And finally, we edit each film subtitled both in English and in Galician and segment it into subtitles.

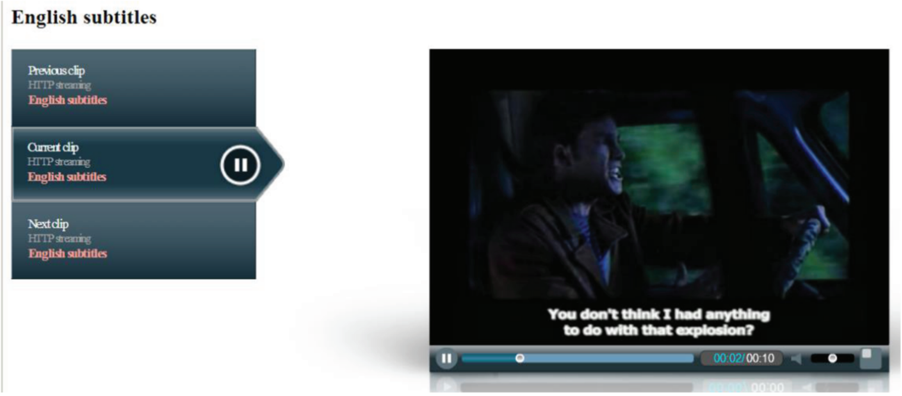
[7] <http://sli.uvigo.gal/CLUVI/vmm.html>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



335-CHU (572)	**You** say ¶ **you**'ll do anything for me, and this is what I get.	E ti dicías que farías calquera cousa por min...	🎞
336-CHU (573)	**You** don't think I had anything ¶ to do with that explosion?	¿Non crerás que teño algo ¶ que ver con esa explosión?	🎞
337-CHU (574)	Didn't **you**?	¿E logo non?	🎞

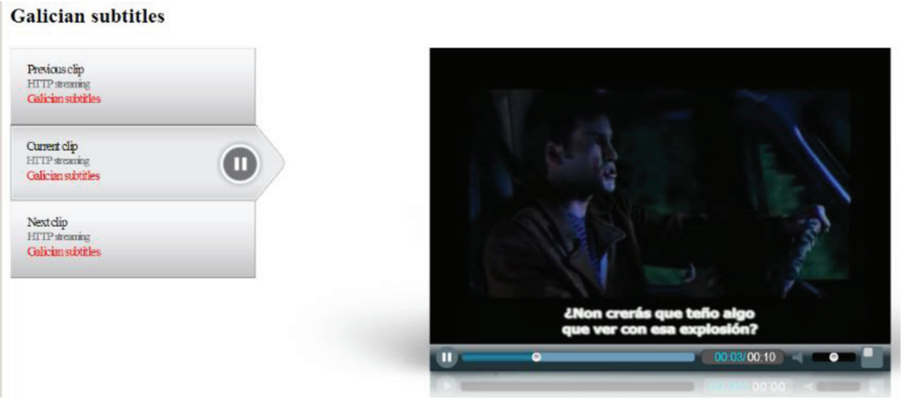
Figure 3. Search results in VEIGA



English subtitles

You don't think I had anything to do with that explosion?

Figure 4. English subtitles in VEIGA



Galician subtitles

¿Non crerás que teño algo que ver con esa explosión?

Figure 5. Galician subtitles in VEIGA

Therefore, we come up with two subsets of subtitled video clips, one in English and the other in Galician, each made up of as many videos as subtitles in the corresponding film. Moreover, given that a high number of subtitles are not long enough to be played and watched properly (they are only one or two seconds long), each individual clip/subtitle is allotted ten extra seconds – five seconds before the subtitle shows up, and five seconds after it fades out –, thus providing the viewer with some context. Once two sets of subtitled clips for each film are obtained, we

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

```
<!-- VEIGA DTD -->
<!ELEMENT cluvi_veiga (header, body) >
<!ATTLIST cluvi_veiga
      version CDATA #REQUIRED >
<!ELEME NT header (#PCDATA)>
<!ELEMENT body (tu*) >
<!ELEMENT tu (tuv+) >
<!ELEMENT tuv (seg) >
<!ATTLIST tuv
      xml:lang CDATA #REQUIRED>
<!ELEMENT seg (#PCDATA | s | l | hi | ph)*>
<!ELEMENT hi (#PCDATA | l)*>
<!ATTLIST hi
      type CDATA #IMPLIED
      x CDATA #IMPLIED>
<!ELEMENT ph EMPTY>
<!ATTLIST ph
      x CDATA #IMPLIED>
<!ELEMENT s EMPTY>
<!ATTLIST s
      n CDATA #IMPLIED
      d CDATA #IMPLIED
      a CDATA #IMPLIED>
<!ELEMENT l EMPTY>
```

Figure 6. VEIGA Corpus XML specification

link them to their corresponding text in the bitextual TMX-based CLUVI representation by means of their video clip identification tag, encoded both in the TMX file and in the video clip (in its file name).

These two sets of subtitled clips are stored as FLV files (because of their compression rate and small file size) in the server file system, where they are named – with a unique file name – according to their film title, their subtitle language (English or Galician), and their sequential number. Thus, whenever users search the VEIGA they get both the bilingual text pair and the clips where this text/subtitle appears. On the other hand, the bitextual TMX files are stored with a file name according to their film title, and include the tags of both the in-cue and out-cue time of each subtitle and their sequential number. This information is encoded in the VEIGA Corpus with a second element added to the tagging, the <*s* > element, which contains three attributes for each tagged subtitle: *s* for the sequential number, *d* for the in-cue time, and *a* for out-cue time. Furthermore, line breaks within subtitles are encoded with the <*l*/> tag, an element added to the TMX 1.4 specification to allow for the examination of aspects which are relevant to subtitling, such as typographical conventions and space constraints. To illustrate this tagging, Figure 7 shows a fragment of the code included in the TMX file named peixe.tmx (from the film entitled *Shooting Fish*, by Stefan Schwartz)

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

```
      <tu>
       <tuv xml:lang=“en”><seg><s n=“848” d=“01:07:32,351”
a=“01:07:34,342 ”/>We play our cards right,<l/>we could end up with… <s
n=“849”d=“01:07:34,431 ” a=“01:07:36,023 ”/>two million pounds of
tobacco<l/>to spend it for us.</seg></tuv>
       <tuv xml:lang= “gl”><seg><s n= “808” d=“01:07:33,271 ”
a=“01:07:36,946 ”/>Podemos gastar 2 millóns en tabaco<l/>e pósters de Pamela
Anderson.</seg></tuv>
      </tu>
      <tu>
       <tuv xml:lang= “en”><seg><s n= “850” d=“01:07:36,511 ” a=“01:07:39,025 ”/>I
meant to get someone<l/>to spend it for us.</seg></tuv>
       <tuv xml:lang= “gl”><seg><s n= “809” d=“01:07:37,071 ”
a=“01:07:39,539 ”/>Buscaremos alguén<l/>que o gaste por nós.</seg></tuv>
      </tu>
```

Figure 7. Fragment of the VEIGA Corpus

that would correspond to the video clips stored in the file system as peixe_en-848.flv, peixe_en-849.flv, peixe_gl-808.flv; and peixe_en-850.flv, and peixe_gl-809.flv.

With regard to its applications, the VEIGA Corpus may serve a number of potential uses and purposes. First, it may be exploited as a reservoir of examples, offering researchers and scholars a database to analyse the different strategies and procedures used in both interlingual and intralingual subtitling and helping them substantiate their theoretical assumptions with practical evidence. From a pedagogical perspective, the VEIGA features suggest that it could be used for different purposes in various learning settings, ranging from general language courses dealing with pronunciation, register, collocations, and other features of oral and written discourse, to specialized courses in audiovisual translation with a focus on interlingual and intralingual subtitling (Sotelo Dios 2015; Sotelo Dios 2016). Concerning language learning, the use of assorted “real” texts, and particularly intralingual subtitles for L1 learning and interlingual subtitles for L2 learning, is likely to increase students' motivation and cultural awareness, although careful selection, adaptation and designing of teaching materials and activities coupled with adequate teacher guidance need to be in place. At the same time, the VEIGA multimedia corpus may also prove a useful e-learning tool, since it would provide students with the possibility of exploring textual properties while listening to and watching film clips, which can be played and stopped at will, thus promoting autonomous learning. Finally, professional subtitlers could also benefit from the possibility to access a collection of ready-made subtitles, where they can look at how other practitioners solved particular subtitling challenges.

3. The SensoGal Corpus

This section will review the different methodologies and resources used to build the SensoGal Corpus,[8] an English–Galician parallel corpus semantically annotated with respect to WordNet and aligned at sentence and word level. The original English texts included in the SensoGal Corpus come from the SemCor Corpus, a textual corpus semantically annotated at lexical level and formed by 360,000 words distributed among 352 texts taken from the Brown Corpus (Landes/Leacock & Tengi 1998). The words of SemCor are tagged with an indication of the particular sense that they possess in their context of occurrence. This tagging uses the senses established in the English WordNet, a lexical resource elaborated by the same research team from the University of Princeton who carried out the annotation of the SemCor corpus. This is currently the largest semantically annotated and freely available corpus of real texts of a language, with 192,639 content words (nouns, verbs, adjectives and adverbs) annotated with their sense with respect to WordNet.

WordNet is a lexical database of the English language, organized as a semantic network where the nodes are concepts represented as sets of synonyms and the links between nodes are semantic relations between lexical concepts (Miller et al. 1990). The nodes contain nouns, verbs, adjectives and adverbs grouped by synonymy. In WordNet terminology, a set of synonyms is called a synset. Thus, each synset represents a distinct lexicalized concept and includes all the synonymous variants of this concept. In the WordNet model of lexical representation, the synsets are linked by means of lexical-semantic relations. Some of the most frequent relations represented in WordNet are hypernymy/hyponymy and holonymy/meronymy for nouns; antonymy and quasi-synonymy for adjectives; antonymy and derivation for adverbs; and entailment, hypernymy/hyponymy, cause and opposition for verbs.

WordNet, which was originally developed for English, is now available in many other languages, although the English WordNet still stands as the most complete reference version. Created and maintained at Princeton University since 1985, version 3.0 – used in the SensoGal Corpus – contains 206,941 lemmas, that is synonymous variants (155,287 of which are unique, non-homographic forms) grouped into 117,659 sets of synonyms or synsets. Many of the WordNet versions in languages other than English follow the design model of EuroWordNet (Vossen 2002), where the synsets of a particular language are linked to the synsets of the other languages through an InterLingual Index (ILI) that is unique to each concept, and which is mainly based on the synsets of the English WordNet. Therefore, the set of WordNet lexicons in different languages allows the connection between the synsets of any pair of languages via the ILI index, thus constituting a very

[8] ISLRN: 653–144–288-768-2.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

useful resource in applications of linguistic technologies dealing with multilingual processing such as SensoGal.

The semantic tagging in the SensoGal Corpus is based on English WordNet 3.0 and on Galnet,[9] the Galician WordNet. The goal of the Galnet project (Gómez Clemente et al. 2013; Solla Portela & Gómez Guinovart 2015; Álvarez de la Granja/ Gómez Clemente & Gómez Guinovart 2016), carried out at the SLI, is building a WordNet for Galician aligned with the English WordNet 3.0, following the expand model (Vossen 1998) for the creation of new wordnets, where the variants associated with the Princeton WordNet synsets are translated using different strategies. Table 2 shows the lexical coverage of English WordNet 3.0, and the current status for Galician in its development version 3.0.25 as available via Galnet's web interface.

Table 2. WordNet synsets and variants by language

	English (WordNet 3.0)		**Galician (Galnet 3.0.25)**	
	variants	synsets	variants	synsets
Nouns	146,312	82,115	45,810	30,565
Verbs	25,047	13,767	6,985	3,069
Adjectives	30,002	18,156	10,317	6,312
Adverbs	5,580	3,621	1,643	1,074
Total	206,941	117,659	64,755	41,020

The process of creation of the SensoGal Corpus begins with the adaptation to WordNet 3.0 of the semantic tagging of the SemCor Corpus, originally annotated with respect to WordNet 1.6. Then, the manual translation of the texts into Galician is carried out and, simultaneously, the new variants derived from the translation are introduced into the Galician WordNet. After the translation, the semantic labels of English are projected on to the Galician texts. Finally, a parallel English–Galician corpus is built up in TMX with the results of the semantic annotation of Galician texts (Solla Portela & Gómez Guinovart 2017).

The XML specification for SensoGal encoding follows the general conventions of the TMX format, as shown in the fragment of the corpus in Figure 8. The tagged versions of the original and translated sentences are stored as specialized variants of the translation unit (<*tu* > element) as character data (*CDATA*), in order to distinguish between the TMX structural markup and the internal sentence labelling. All content words in a sentence (nouns, verbs, adverbs and adjectives) are annotated with their lemma and tagged with their sense expressed as their

[9] ISLRN: 544–286–653-437-9.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

```
    <tu>
    <prop type=“group”>br-g15:41</prop>
    <tuv xml:lang=“en”>
    <seg>This man's isolation is not merely momentary, it is permanent.</seg>
    </tuv>
    <tuv xml:lang=“en-tag”>
    <seg>
    <![CDATA[<wf pos= “DT”>This</wf> <wf  lemma=“man” ili=“ili-30-
10287213 -n”>man</wf> <wf pos=“POS”>'s</wf> <wf lemma= “isolation” ili=“ili-
30-14414715 -n”>isolation</wf> <wf lemma= “be” ili=“ili-30-02604760 -
v”>is</wf> <wf lem ma=“not” ili=“ili-30-00024073 -r”>not</wf> <wf
lemma=“merely” ili=“ili-30-00004722 -r”>merely</wf> <wf lemma= “momentary”
ili=“ili-30-01443097 -a”>momentary</wf> <punc>,</punc> <wf
pos=“PRP”>it</wf> <wf lemma= “be” ili=“ili-30-02604760 -v”>is</wf> <wf
lemma=“permanent” ili=“ili-30-01754421 -a”>permanent</wf>
<punc>.</punc>]]>
    </seg>
    </tuv>
    <tuv xml:lang=“gl”>
    <seg>Este illamento do home non é só momentáneo, é permanente.</seg>
    </tuv>
    <tuv xml:lang=“gl-tag”>
    <seg>
    <![CDATA[Este <wf lemma= “illamento” ili=“ili-30-14414715 -
n”>illamento</wf> do <wf lemma= “home” ili=“ili-30-10287213 -n”>home</wf>
<wf lemma=“non” ili=“ili-30-00024073 -r”>non</wf> <wf lemma=“ser” ili=“ili-
30-02604760 -v”>é</wf> <wf lemma= “só” ili=“ili-30-00004722 -r”>só</wf> <wf
lemma=“momentáneo” ili=“ili-30-01443097 -a”>momentáneo</wf>, <wf
lemma=“ser” ili=“ili-30-02604760 -v”>é</wf> <wf lemma= “permanente” ili=“ili-
30-01754421 -a”>permanente</wf>.]]>
    </seg>
    </tuv>
    </tu>
```

Figure 8. Fragment of the SensoGal Corpus

WordNet ILI index. In addition, the <*prop* > element, used in TMX “to define the various properties of the parent element” (Savourel 2005), is used in the SensoGal encoding to identify all the phrases in the corpus with their bibliographical reference (abbreviation of text title and phrase number).

So far thirty Galician translations have been semantically tagged and aligned with their corresponding original English texts, totalling 2,734 translation units with 61,236 English words and 62,577 Galician words. The resulting parallel corpus can be accessed for consultation through a dedicated web interface.[10]

[10] Available at <http://sli.uvigo.gal/SensoGal/>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. Conclusion

Since its publication, the CLUVI Corpus has been used as the empirical basis for a wide range of academic studies in the fields of translational stylistics (Moreira 2010; Moreira 2011a; Sotelo Dios 2011), translation teaching (Sotelo Dios 2015; Sotelo Dios 2016), computational lexicology (Girju 2007a; Girju 2007b; Gómez Guinovart & Oliver 2014), terminology (Gómez Guinovart & Torres Padín 2006; Crespo et al. 2008; Gómez Guinovart & Simões 2009; Simões & Gómez Guinovart 2009; Moreira 2011b; Gómez Guinovart 2012; Moreira 2014) and multilingual lexicography (Gómez Guinovart & Sacau Fontenla 2004a; Gómez Guinovart & Sacau Fontenla 2005; Gómez Guinovart/Díaz Rodríguez & Álvarez Lugrís 2008; Gómez Guinovart & Simões 2010; Álvarez Lugrís & Gómez Guinovart 2014).

More generally, over the last decade, parallel corpora have proven very useful in many different applications in the fields of translation and language teaching, computational terminology and lexicography, applied linguistics, computer-aided translation and machine translation. The incorporation of multimedia data into parallel corpora will permit to transcend the traditional text-only approach to corpus design, reflecting the polisemiotic aspects of cultural products like film discourse and subtitling. Without any doubt, multimedia parallel corpora will represent in the very next future an important resource in the areas of language and cultural studies, second and foreign language teaching and in translation studies and professional practice.

Furthermore, the annotation of lexical meaning in parallel corpora increases their possibilities of exploitation in applied linguistics and language processing. Lexical-semantic processing is crucial for information society key technologies like conversational agents, information extraction and question-answering systems. Although much effort is still required to complete this task, the SemCor Corpus certainly represents a resource of vital importance for the development of Galician language technologies. Its exploitation should enable the construction of tools of great interest in the field of semantic processing – especially in tasks that require multilingual knowledge –, and the development of more efficient applications for language processing.

References

Almeida, José João, Araújo, Sílvia, Simões, Alberto & Dias, Idalete. 2014. The Per-Fide Corpus: A New Resource for Corpus-based Terminology, Contrastive Linguistics and Translation Studies. In *Working with Portuguese Corpora*, Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds), 177–200. London: Bloomsbury Publishing.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Álvarez de la Granja, María, Gómez Clemente, Xosé María & Gómez Guinovart, Xavier. 2016. Introducing idioms in the Galician wordnet: methods, problems and results. *Open Linguistics* 2: 253–286.

Álvarez Lugrís, Alberto & Gómez Guinovart Xavier. 2014. Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del Dicionario Moderno Inglés-Galego. In *Lexicografía de las lenguas románicas: Aproximaciones a la lexicografía moderna y contrastiva*, María José Domínguez Vázquez, Xavier Gómez Guinovart Xavier & Valcárcel Riveiro Carlos (eds), 31–48. Berlin/Boston: De Gruyter Mouton.

Crespo Bastos, Ana, Gómez Clemente, Xosé María, Gómez Guinovart Xavier & López Fernández Susana. 2008. XML-based Extraction of Terminological Information from Corpora. In *Actas da 6ª Conferência Nacional XATA2008: XML, Aplicações e Tecnologias Associadas*, José Carlos Ramalho, João Correia Lopes & Salvador Abreu (eds), 28–39. Évora: Universidade de Évora.

Girju, Roxana. 2007a. Experiments with an Annotation Scheme for a Knowledge-rich Noun Phrase Interpretation System. In *Proceedings of the Linguistic Annotation Workshop*, 168–175. Prague: ACL.

Girju, Roxana. 2007b. Improving the Interpretation of Noun Phrases with Cross-linguistic Information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 568–575. Prague: ACL.

Gómez Clemente, Xosé María, Gómez Guinovart, Xavier, González Pereira, Andrea & Verónica Taboada Lorenzo. 2013. Sinonimia e rexistros na construción do WordNet do galego. *Estudos de lingüística galega* 5: 27–42.

Gómez Guinovart Xavier & Oliver, Antoni. 2014. Methodology and evaluation of the Galician WordNet expansion with the WN-toolkit. *Procesamiento del Lenguaje Natural* 53: 43–50.

Gómez Guinovart Xavier & Sacau Fontenla Elena. 2004a. Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del Lenguaje Natural* 33: 133–140.

Gómez Guinovart Xavier & Sacau Fontenla Elena. 2004b. Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds), 1179–1182. Paris: ELRA.

Gómez Guinovart Xavier & Sacau Fontenla Elena. 2005. Técnicas para o desenvolvemento de dicionarios de tradución a partir de córpora aplicadas na xeración do Dicionario CLUVI Inglés-Galego. *In Viceversa* 11: 159–171.

Gómez Guinovart Xavier & Simões, Alberto. 2009. Parallel corpus-based bilingual terminology extraction. In *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence*. Toulouse: Université Paul Sabatier. <https://www.irit.fr/TIA09/thekey/posters/simoes-guinovart.pdf> (28 April 2017).

Gómez Guinovart, Xavier & Simões, Alberto. 2010. Translation dictionaries triangulation. In *Proceedings of FALA2010: VI Jornadas en Tecnología del Habla & II Iberian SLTech*, Carmen García Mateo, Francisco Campillo Díaz & Francisco Méndez Pazó (eds), 171–174. Vigo: Universidade de Vigo.

Gómez Guinovart, Xavier & Torres Padín, Ánxeles. 2006. Extracción dun vocabulario xurídico-administrativo galego-castelán a partir dun corpus paralelo. In *Terminología y derecho: la complejidad de la comunicación multilingüe*, M. Teresa Cabré, Carme Bach & Jaume Martí (eds), 175–188. Barcelona: Universitat Pompeu Fabra.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Gómez Guinovart, Xavier, Díaz Rodríguez, Eva & Álvarez Lugrís, Alberto. 2008. Aplicacións da lexicografía bilingüe baseada en córpora na elaboración do Dicionario CLUVI inglés-galego. *Viceversa* 14: 71–87.

Gómez Guinovart, Xavier. 2012. A hybrid corpus-based approach to bilingual terminology extraction. *In Encoding the Past, Decoding the Future: Corpora in the 21st Century*, Isabel Moskowich-Spiegel Fandiño & Begoña Crespo (eds), 147–175. Newcastle upon Tyne: Cambridge Scholar Publishing.

Keshtkar, Hossein & Mosavi Miangah, Tayebeh. 2012. Using Bilingual Parallel Corpora in Translation Memory Systems. *International Journal of Applied Linguistics and English Literature* 1.5: 184–193.

Koehn, Philipp. 2005. EuroParl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X: The Tenth Machine Translation Summit Proceedings*, 79–86. Tokyo: Asia-Pacific Association for Machine Translation.

Landes, Shari, Leacock, Claudia & Tengi, Randee I. 1998. Building semantic concordances. In *WordNet: An Electronic Lexical Database*, Christiane Fellbaum (ed), 199–216. Cambridge: The MIT Press.

Mikhailov, Mikhail & Cooper, Robert. 2016. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. Abingdon: Routledge.

Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek & Miller, Katherine. 1990. WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3: 235–244.

Montero Perez, Maribel, Paulussen, Hans Macken, Lieve & Desmet, Piet. 2014. From input to output: the potential of parallel corpora for CALL. *Language Resources and Evaluation* 48.1: 165–189.

Moreira, Adonay. 2010. Estratégias de tradução em sites das regiões de turismo de Portugal: estudo baseado em corpus. *Polissema: Revista de Letras do ISCAP* 10: 13–42.

Moreira, Adonay. 2011a. The translator as cultural mediator: a corpus-based study of omissions and additions in translations of tourism brochures. *The Journal of Cultural Mediation* 1: 86–95.

Moreira, Adonay. 2011b. Turigal: compilation of a parallel corpus for bilingual terminology extraction. In *Actas del III Congreso Internacional de Lingüística de Corpus: Las tecnologías de la información y las comunicaciones: presente y futuro en el análisis de corpus*, María Luisa Carrió & Miguel Ángel Candel (eds), 33–42. València: Universitat Politècnica de València.

Moreira, Adonay. 2014. A methodology for building a translator- and translation-oriented terminological resource. In *inTRAlinea Special Issue: Translation & Lexicography*, María Sánchez, María Porciel & Iris Serrat (eds). < http://www.intralinea.org/specials/article/2032 > (28 April 2017).

Santos, Diana. 2004. *Translation-based Corpus Studies: Contrasting English and Portuguese Tense and Aspect Systems*. Amsterdam: Rodopi.

Savourel, Yves. 2005. *TMX 1.4b Specification*. Localisation Industry Standards Association. <https://www.gala-global.org/tmx-14b> (28 April 2017).

Simões, Alberto & Gómez Guinovart, Xavier. 2009. Terminology extraction from English–Portuguese and English–Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns. In *Proceedings of the Iberian SLTech 2009 - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, António Teixeira, Miguel Sales Dias & Daniela Braga (eds), 13–16. Porto Salvo: Designeed.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Simões, Alberto, Gómez Guinovart, Xavier & Almeida, José João. 2004. Distributed translation memories implementation using WebServices. *Procesamiento del Lenguaje Natural* 33: 89–94.

Solla Portela, Miguel Anxo & Gómez Guinovart, Xavier. 2015. Galnet: o WordNet do galego. *Aplicacións lexicolóxicas e terminolóxicas. Revista Galega de Filoloxía* 16: 169–201.

Solla Portela, Miguel Anxo & Gómez Guinovart, Xavier. 2017. Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0. *Procesamiento del Lenguaje Natural* 59: 137–140.

Sotelo Dios Patricia & Guinovart Xavier, Gómez. 2012. A multimedia parallel corpus of English–Galician film subtitling. In *1st Symposium on Languages, Applications and Technologies*, Alberto Simões, Ricardo Queirós & Daniela da Cruz (eds), 255–266. Saarbrücken: Dagstuhl Publishing.

Sotelo Dios, Patricia. 2011. Using a multimedia parallel corpus to investigate English–Galician subtitling. In *Proceedings of the SDH 2011 Conference: Supporting Digital Humanities*, Bente Maegaard (ed). Copenhagen: University of Copenhagen. <http://hnk.ffzg.hr/bibl/SDH-2011/proceedings.html> (28 April 2017).

Sotelo Dios, Patricia. 2015. Using a multimedia corpus of subtitles in translation training. In *Affordances of Language Corpora for Data-driven Learning*, Agnieszka Leńko-Szymańska & Alex Boulton (eds), 245–266. Amsterdam: John Benjamins.

Sotelo Dios, Patricia. 2016. Adquisición de competencias en traducción audiovisual mediante un corpus multimedia. In *New Insigths into Corpora and Translation*, Daniel Gallego Hernández (ed), 1–16. Newcastle upon Tyne: Cambridge Scholars Publishing.

Tiedemann, Jörg. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Nicoletta Calzolari, Khalid Choukri,Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 2214–2218. Istanbul: ELRA.

Tufiş, Dan. 2007. Exploiting Aligned Parallel Corpora in Multilingual Studies and Applications. In *Intercultural Collaboration*, Toru Ishida, Susan R. Fussell & Peek Vossen (eds), 103–117. Berlin: Springer.

Véronis, Jean, ed. 2000. *Parallel Text Processing: Alignement and Use of Translation Corpora*. Dordrecht: Kluwer.

Vossen, Piek. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Norwell: Kluwer Academic Publishers.

Vossen, Piek. 2002. WordNet, EuroWordNet and Global WordNet. *Revue française de linguistique appliquée* 7: 27–38.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Discourse annotation in the MULTINOT corpus

Issues and challenges

Julia Lavid López
Complutense University of Madrid

This chapter summarises and discusses recent work on the development of a bilingual (English-Spanish) corpus consisting of original comparable and parallel texts from a variety of genres and annotated with complex linguistic features such as modality and evidentiality, metadiscourse markers, and thematization, as carried out within the framework of the MULTINOT project. The annotation of these complex features in bilingual parallel texts poses important challenges for the researcher at the different stages of the corpus development, from the preprocessing phases to the manual annotation phase, but, at the same time, it allows the investigation of complex linguistic research questions which could not be addressed on the basis of raw corpora or even with the help of an automatic part-of-speech tagging system.

Keywords: discourse, corpus, annotation, English, Spanish

1. Introduction

The current decade has witnessed an increasing demand for more in-depth Natural Language Processing (NLP) research and more powerful applications, which require a greater richness in annotation. In addition, the development of parallel and comparable corpora in different languages for multilingual applications has also led to a demand for greater richness and multiple levels of annotation (see Zeyrek et al. 2013). Among these levels, the annotation of discourse features stands out as a prominent one in recent years, with an increasing number of projects focusing on the annotation of phenomena such as thematization (Lavid et al. 2013), modality and evidentiality in discourse (Lavid et al. 2016a, 2016b), metadiscourse and other discourse markers (Lavid & Moratón 2016; Correia 2016), rhetorical

https://doi.org/10.1075/scl.90.10lop
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

relations (Trnavac et al. 2016), appraisal and opinion (Taboada 2016; Mora 2017), among others.

In this chapter, I describe a number of issues and problems which have emerged in the annotation of discourse phenomena in the bilingual MULTINOT corpus, a high-quality, register-diversified parallel and medium-sized corpus for the English–Spanish pair, consisting of originals and translated texts in both directions and enriched with linguistic annotations, which can be exploited in a number of linguistic, applied and computational contexts. The creation of such a corpus has been carried out within the framework of the MULTINOT project, a research effort jointly developed between two European research groups (FUNCAP at Universidad Complutense Madrid and LT3 at Ghent University) with international expertise in contrastive, corpus-based linguistic and computational investigations.[1]

The chapter is organized as follows: Section 2 describes the MULTINOT corpus; Section 4 describes three main discourse annotation tasks carried out within the project, namely, thematization, modality and evidentiality and metadiscourse markers; Section 5 discusses the main issues and challenges that emerged during the annotation of these discourse phenomena in the bilingual corpus; Section 5 provides a summary and some concluding remarks.

2. The MULTINOT corpus

The MULTINOT corpus is a half-million-word, sentence-aligned, richly-annotated parallel corpus for the language pair English–Spanish. As the MULTINOT corpus is bidirectional, it can be used as a comparable corpus to study contrastive differences between original texts in English and Spanish, between translations in both directions, and to study differences between translated versus non-translated texts in both languages. It distinguishes itself from other parallel corpora by having a balanced composition (both in terms of diversity of registers, genres, and translation directions) and by focusing on quality rather than quantity. With respect to the diversity of registers and genres, the corpus contains five main general text types or domains, which are subdivided into more delicate categories.

[1] The MULTINOT project is financed by the Spanish Ministry of Economy and Competitiveness under project grant FFI2012–32201. As principal investigator, I gratefully acknowledge the support provided by the funding authorities. My special thanks go to members of the UCM and the Ghent teams for their collaboration in the different compilation and annotation tasks of the project. These include Jorge Arús, Marta Carretero, Carmen Maíz, Lara Moratón, Elena Martínez, Juan Rafael Zamorano, Véronique Hoste, Bernard DeClerck and Patrick Goethals. Information about the project and the corpus is available at <https://www.ucm.es/funcap/multinot>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

The five macro-registers or domains are: literature, journalism, instructive texts, administrative texts and external communication, each containing several basic categories, as specified below:

- novels and short stories (from the domain of fiction literature)
- essays and expository popular science texts (from the domain of nonfiction)
- news reporting articles (from the domain of journalism)
- manuals and legal documents from webpages (from the instructive domain)
- official speeches and the proceedings of parliamentary debates (from the administrative domain)
- annual reports and letters of self-presentation from companies, promotion and advertizing brochures, and scientific texts (from the domain of external communication)

Translation direction is an important balancing criterion and the target figure of 50,000 words per translation direction was set for this purpose. Also, in order to

Table 1. Word count distribution of different registers in MULTINOT

MacroRegister	Sub-register	Source => Target	English	Spanish	Total
Literature	Novels	EN => ES	24886	26927	51813
		ES => EN	27939	26672	54611
	Short stories	EN => ES	2186	2088	4274
		ES => EN	1175	1197	2372
	Essays	EN => ES	27382	27235	54617
		ES => EN	32517	30362	62879
Journalism	News reporting articles (popsci)	EN => ES	10658	9753	20411
		ES => EN	24579	23730	48309
Administrative	Official speeches	EN => ES	25373	27112	52485
	Proceedings of debates	EN => ES	25620	26450	52070
		ES = > EN	27390	26320	53710
External communication	Promotion/advertizing brochures	EN = > ES	23695	27790	51485
		ES = > EN	25761	25367	51128
		ES = > EN	31188	27326	58514
	Scientific texts	EN = > ES	42580	45047	87627
		ES = > EN	24322	23430	47742
	Legal information	EN = > ES	36688	38640	75328
		ES = > EN	23984	22185	46169

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

preserve balance in terms of providers, the combination of register and translation direction comes from at least three different providers. Exceptions had to be made to this global design when it was not possible to find material in both translation directions, or when it was difficult to find information on the translation direction. The distribution of registers and number of words in the MULTINOT corpus is graphically displayed in Table 1.

In order to enrich the compiled texts discourse information, a number of annotation layers were added to the MULTINOT corpus:

1. A first layer of automatic preprocessing for "lower" levels of linguistic processing (i.e.: tokenization, segmentation, part-of-speech tagging) using the LeTs Processing Pipeline (Van de Kauter et al. 2013).
2. A second layer of manual and semi-automatic annotation for "higher" levels of linguistic processing, including discourse phenomena such as thematization, modality, and evidentiality and metadiscourse markers. The annotation procedure and the challenges encountered are described in detail in the sections below.

3. Annotation procedure

The annotation of semantic, pragmatic, and discourse phenomena is a complex task: to the general ambiguity in the definition of categories, language-specific problems emerge when annotating bilingual or multilingual texts. It is, therefore, necessary, to develop a scientific annotation procedure to ensure the quality of the annotated data sets.

The annotation procedure used in MULTINOT for the above-mentioned discourse categories follows the methodology proposed in Hovy and Lavid (2010). It consists of six main steps, which are described in the subsections below.

3.1 Selecting the "training" corpus

The first step consisted of compiling a small "training corpora" on which to perform human-coded annotations for the different discourse phenomena of thematization, metadiscourse markers, and modality. In the case of thematization and metadiscourse markers, a bilingual sample of eighteen texts was compiled and hand-coded by two independent annotators. In the case of modality, a set of two hundred and fourty sentences was extracted from different genres from the MULTINOT corpus and manually annotated with modal meanings.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

3.2 Instantiating the theory

The second step consisted of defining and delimiting the theoretical categories to be annotated. The main problems which emerged in this phase were: (a) that the categories were often not exhaustive; (b) that the categories were unclear or difficult to define. The solutions adopted in MULTINOT were basically two: (a) one solution was "neutering" the theory, that is using a less refined, more "neutral" set of terms/categories when agreement between annotators was not possible; (b) another solution was reformulating or redefining the categories to account for unforessen cases not previously captured by the theory.

3.3 Designing annotation schemes and guidelines

This is one of the fundamental steps in corpus annotation and one which requires several phases of refinement. It involves instantiating all or part of the features of the selected theoretical model – developing a core and an extended tagset – and developing guidelines to be used as instructions for human coders.[2] In MULTINOT, the annotation schemes and guidelines were modified and perfected during the annotation process, typically whenever new variants, unforeseen by the theory, were encountered.

3.4 Performing annotation experiments

In order to test the reliability of the annotation scheme and guidelines, it is necessary to perform annotation experiments, or agreement studies on some fragment of the training corpus and evaluate their results. In MULTINOT, annotation experiments were performed for the categories of thematization, modality and evidentiality, and metadiscourse markers, as will be described in Section 5. The procedure involves two or three human coders working independently, handling easier annotations first and acquainting themselves with the data and the task before tackling harder cases. Annotation disagreements were brought to open discussion and jointly resolved ("reconciliation" stage). Backing off occurred in cases of disagreement and continued with several options: (1) making the option granularity coarser (neutering); (2) allowing multiple options; (3) increasing the context supporting the annotation decision.

[2] The theoretical models for each of the annotated categories is outlined in Sections 5, 6 and 7 below.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

3.5 Evaluating the annotations

Once the annotation experiments were performed by independent coders, the next step was to measure their reliability through different types of agreement measures. Intra-annotator agreement measures the stability of the annotations, while inter-annotator agreement measures their reproducibility.

Agreement metrics used in MULTINOT were the Kappa coefficient (Cohen 1960), Krippendorff's alpha (Krippendorff 2004), AGR, Precision and Recall, depending on the type of annotation task and the number of human coders.

3.6 Large-scale annotation of the whole corpus

The final step was the large-scale annotation of the larger MULTINOT corpus with tested categories that were validated through the different agreement studies. In the project, the categories of thematization, modality, and metadiscourse markers were validated empirically through corpus annotation and were annotated in a large proportion of bilingual texts, as described in previous studies by the author of this chapter and other project members (see Lavid & Moratón 2015, 2016; Lavid et al. 2016a, 2016b; Arús et al. 2012).

In the following sections, I will provide a summary of the annotation procedure and results of the annotation of each of these discourse categories: thematization, modality, and metadiscourse markers.

4. Annotating thematization in English and Spanish

For the category of thematization, two annotation schemes were developed, one for English and one for Spanish, with tags based on Lavid's model of thematization (see Lavid et al. 2010: 294–306). The schemes included both coarse-grained and more fine-grained tags, reflecting the range of possible thematic types which can occur as part of the thematic field in English and Spanish declarative clauses. The core tagsets for English and Spanish with illustrative examples are graphically displayed in Tables 2 and 3:

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 2. Core tagset for theme types in English

Annotation layer	Thematic field	Description	Realizations
Unit		Main clause	
Core annotation scheme			
Tags:	TH (Thematic Head)	First nuclear constituent (Participant or Process, not Circumstantial) in main clause	*The cat is on the mat* *Eating is vital* *Eat your soup!* *On the table stood a lamp*
	PreHead (PH)	Any Circumstantial element and/or Finite element preceding the Thematic Head.	– *Adverbial Groups (e.g. [PH-Circ:] Afterwards there will be another meeting)* – *Prepositional Phrases (e.g. [PH-Circ:] On your right you can see the Royal Palace)* – *Circumstantial clauses (e.g. [PH-CCL:] After dropping her off, he continued his trip)* – *Finite verbal forms, that is auxiliaries, preceding the lexical verb: (e.g. [PH-Finite:] Should you decide to leave the country, please let me know. Had I known you were so near, …*
	Textual Theme (TT)	Elements which are instrumental in the creation of the logical connections in the text, such as linkers, binders and other textual markers.	*Linkers (paratactic nexus) (e.g. [TT-Link:] And don't tell me you didn't know; but let's change the topic).* – *Binders (hypotactic nexus) (e.g. [TT-Bind:] However, the situation now is different;* – *Correlatives: (not only …but; either …or) (e.g. [TT-Cor:] Not only didn't he call but he also forgot about us completely.*

(*continued*)

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 2. (*continued*)

Annotation layer	Thematic field	Description	Realizations
	Interpersonal Theme (IT)	Elements which express the attitude and the evaluation of the speaker with respect to his/her message, such as Vocatives and Modal Adjuncts, including mood and comment adjuncts	– *Vocatives, that is, any item used to address (e.g. [IT-Voc:] Tom! This is a nice surprise; Sir, could you follow me, please?).* – *Comment Adjuncts (e.g. [IT- Com:] Surprisingly he didn't mention anything;* – *Modal Adjuncts (e.g. [IT- Mod:] Probably that's the only lesson we learned; Surely you didn't do that!).*
	Predicated Theme (PT)	Construction that consists of two parts: (1) an initial thematic segment consisting of "It" + BE followed by the element in Focus; and (2) a rhematic segment realized by a relative-like clause	*It is you who are to blame*
	"There" type construction		*There were three people waiting for the bus*

Table 3. Core tagset for theme types in Spanish

Annotation layer	Thematic field	Description	Realizations
Unit		Main clause	
Core annotation scheme			
Tags:	TH (Thematic Head)	First nuclear element (not circumstantial) in main clause, realized by either lexical or morphological means.	*El gato está en la alfombra; Se está muy bien aquí; Corriendo no se consigue nada;* *que me digas eso significa que no me has entendido; Teng-o frío; ayer v-i a María;* *Ten cuidado!*

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 3. (*continued*)

Annotation layer	Thematic field	Description	Realizations
	PreHead (PH)	Elements preceding the Thematic Head, including: Circumstantials, pronominal 'se', lexical part of Verbal Group.	*Adverbial Groups (e.g. [PH-Circ:] Mañana nos vemos).* – *Prepositional Phrases (e.g. [PH-Circ:] En tal caso, será mejor no hacer nada).* – *Circumstantial clauses (e.g. [PH-Circ:] Sin mediar palabra, le dio una bofetada).* *Middle marking (me, te, se, etc.) realizations:* – *Personal pronouns (not reflexive but morphologically identical to these); (e.g. [PH-middle marking:] Se me cayó, ¿te convences?, Nos fuimos pronto).* – *Lexical part of Process (Verbal Group minus inflectional ending) realizations:* – *Predicator minus inflectional ending (e.g. [PH-lexical part:] Teng-o frío; v-i a María con su novio* – *Finite minus inflectional ending (e.g. [PH-lexical part:] H-e comido demasiado; Esta-mos hartos*
	Textual Theme(TT)	Elements which are instrumental in the creation of the logical connections in the text, such as linkers, binders, and other textual markers.	– *Linkers (paratactic nexus) (e.g. [TT-Link:] ¿O te crees más listo que los demás?; pero, bueno,vamos a dejarlo).* – *Binders (hypotactic nexus) (e.g. [TT-Bind:] Además, tú no sabes nada de mí; por lo tanto, nos vimos obligados a cerrar).* – *Correlatives: (no solo …sino que; o …o) (e.g. [TT-Cor:] No solo nos toma por tontos …*

(*continued*)

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 3. (*continued*)

Annotation layer	Thematic field	Description	Realizations
	Interpersonal Theme (IT)	Elements which express the attitude and the evaluation of the speaker with respect to his/her message.	– *Vocatives, that is, any item used to address (e.g. [IT- Voc:] ¡Profesor! ¿puedo hablar un momento con usted?; tío, esto es la bomba).* – *Comment Adjuncts (e.g. [IT- Com:] desgraciadamente no podremos acudir a tu*! *esta).* – *Modal Adjuncts (e.g. [IT- Mod:] Tal vez esté en su casa;*
	Predicated Theme (PT)	A construction that consists of two parts: (1) an initial thematic segment consisting of the copularverb followed by the element in focus; and (2) a rhematic segment realized by a relative-like clause.	*Fue Fermín el que me dejó triste*
	"Hay" type construction	Construction which occurs in existential clauses, realized by the "Hay" element and its temporal variants "había", "hubo", "habrá" followed by the Existent.	*Había tres chicas esperando en la puerta*

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

On the basis of these core tagsets, two agreement studies were carried out to test the reliability of the core and the extended tagsets:

1. the first agreement study measured inter-annotator agreement on the *identification* of thematic spans;
2. the second agreement study measured inter-annotator agreement on the *type of label* chosen by the annotators on the previously selected spans.

The agreement metric (AGR) was used for the first agreement study because the annotators could be coding different expressions (markables) in identifying thematic spans. For the second agreement study, *kappa coefficient* (K) was used, since it measures agreement when two independent coders are analysing the same element.

The results of the annotation showed that the task of identifying thematic spans proved to be more difficult than that of labeling (e.g. AGR for IT was only of 37.50%) for the English data set (see Table 4):

Table 4. Summary of agreement study 1

	English	**Spanish**
Task 1: Identification of Thematic field	AGR = 97%	AGR = 91%
Task 2: Identification of thematic spans realizing core tags	AGR for TH = 93.84% AGR for TT = 96.50% AGR for PH = 78.70% AGR for IT = 37.50%	AGR for TH = 91%AGR for TT = 97% AGR for PH = 89% AGR for IT = 94%

Labeling of specific Thematic Head types was more difficult for the Spanish data set than for the English one, as shown by the lower agreement (kappa = 0.475) in Table 5.

Table 5. Summary of agreement study 2

	English	**Spanish**
Task 1: Labelling of core tags	Kappa = 0.915	Kappa = 0.839
Task 2: Labelling of Thematic Head types	Kappa = 0.875	Kappa = 0.475

This lower agreement is probably due to the complex morphology of the Spanish verbal group, which caused annotators to hesitate between thematic features such as PreHead or Thematic Head. The inherent difficulties in disambiguating different types could only be improved through consistent training and practice with the annotators. A similar tendency was observed for the annotation of modality, as described below.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

5. Annotating modality in English and Spanish

Annotating modal meaning is a rather difficult task (see Lavid et al. 2016a, 2016b). The difficulties derive not only from the practicalities of the annotation process, but also from the subtle distinctions which emerge in the modal domain. The complexity increases when dealing with more than one language, given the language-specific features that have to be considered in the annotation process. One source of difficulty is the open set of diverse elements that can be triggers of modality in both languages, such as modal verbs (*might, may, must*), adverbials (*perhaps, clearly*), lexical elements (*probable, probability*, etc.), and syntactic constructions (*I think, they say*). In the NLP community, the last years have witnessed the development of annotation schemes and annotated corpora for different aspects of modality in different languages (McShane et al. 2004; Wiebe et al. 2005; Szarvas et al. 2008; Saurí & Pustejovsky 2009; Hendrickx et al. 2012; Baker et al. 2012; Nissim et al. 2013). However, so far there are no annotation schemes for modality that specifically address the commonalities and the language-specific features of English and Spanish, although there are various proposals for English and for other languages. One of the tasks involved in creating MULTINOT was to fill a gap in this area by approaching the annotation of modality from two interrelated perspectives: (a) the functional/semantic perspective, which captures the commonalities shared between both languages in the creation of modal meanings, and (b) the syntactic perspective, which considers the language-specific preferences in the encoding of similar modal meanings in English and Spanish.

The approach is similar to the one proposed by Nissim et al. (2013) in proposing two distinct layers of annotation to capture both the functional/semantic commonalities between English and Spanish and the language-specific syntagmatic realizations. However, it differs from it in several respects: first, it is more fine-grained in terms of the modal meanings included in the annotation scheme; second, the notion of factuality is not included since this is a separate dimension which cuts across the four basic modal meanings but does not serve to establish distinctions between them; third, a highly-detailed and linguistically-sophisticated specification of the syntagmatic options available to English and Spanish for the expression of modal meanings is proposed.

The functional-semantic tagset is graphically displayed in Figure 1 below, and is the result of a number of preliminary annotation experiments during which the tags were elaborated and refined until a consensus was reached on their basic and the secondary meanings. The tagsets are hierarchical, allowing coders to choose the coarser tags from the core tagset (EP, DE, DY, VO), when in doubt about the more fine-grained subtypes from the extended tagset. For example, if the coder is uncertain about whether a markable is “possibility” or “probability”, s/he can

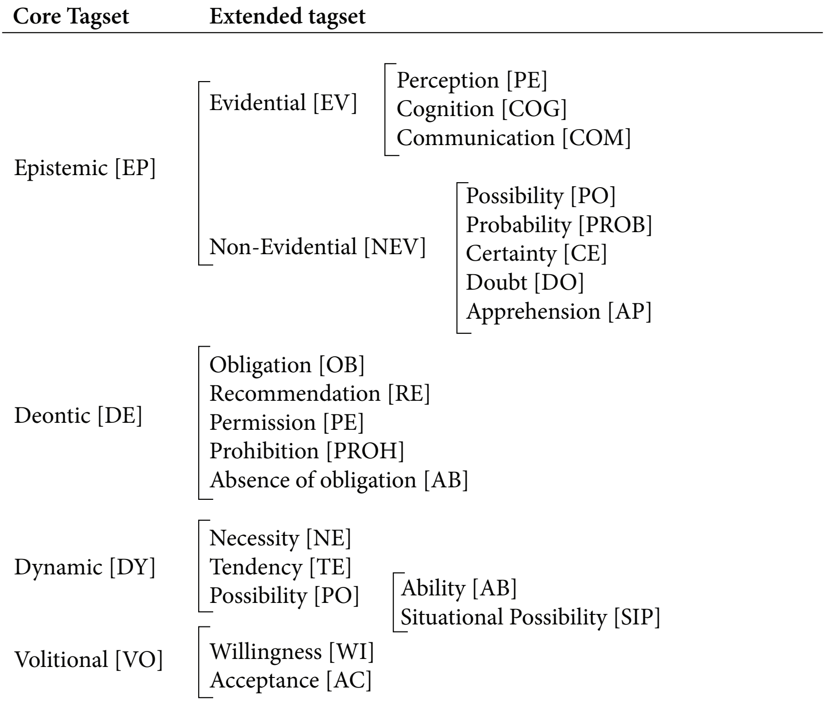


Figure 1. Functional-semantic tagset for modal meanings in English and Spanish (after Lavid et al. 2016a)

simply tag it as "epistemic" [EP] and "non-evidential" [NEV]. The abbreviated form of each tag is given in capital letters in brackets next to the full form.

An in-depth discussion of all these tags and their subtypes is beyond the scope of this chapter.[3] However, an outline with some basic meanings is provided as follows:

The basic tag referring to epistemic meanings is [EP]. These meanings express a qualification of the truth of a proposition (Boye, 2012), and are divided into two main subtypes: (a) *evidential meanings*, defined in terms of the notion of source of information, evidence, or epistemic justification; and (b) *non-evidential meanings*, referring to the degree of certainty or epistemic support.

For evidential meanings, we use the tag [EV], and include all those meanings which qualify the truth of a proposition by expressing the source of the evidence that the speaker has or claims to have at his/her disposal, for or against this truth, as in (1) where the source is perceptual [PE], (2) where the source is cognitive [COG]; or (3) where the source is communicative [COM], referring to evidence coming from linguistic messages:

[3] For a detailed description and discussion of these tags, see Lavid et al. (2016a).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

(1) I **saw** that at this end of the hall there was another chamber, just five or six feet away from where I was standing. (EOFICTION-010)

(2) Martin Rees, Britain's astronomer royal, **believes** that there are many universes, possibly an infinite number (EO-EXPE- 001)

(3) We often **hear** that overseas aid should be a government responsibility, not left to privately run charities. (EO-ESSAY-010)

For non-evidential meanings the tag [NEV] is used, and refers to those meanings concerned with the estimation of the chances of a proposition to be or become true, but not by means of qualifying evidence for or against it. The speaker/writer may express knowledge that the proposition is true or false, or else s/he may not be sure about its truth or falsity and therefore proceed in different ways depending on the degree of certainty conferred to the proposition: (a) express the possibility that the proposition is true (50 percent probability), represented by the tag [POS], as in (4); (b) assign a higher degree of probability, represented with the tag [PROB], as in (5), also including the expression of opinion through mental state predicates such as "*I think*", "*I suppose*", and their Spanish counterparts; and (c) indicate a high degree of commitment to the truth of the proposition (almost or total certainty), represented with the tag [CERT], as in (6):

(4) The altar boy did not move. He was eyeing Langdon closely now. "You look fami liar." Teabing huffed. "**Perhaps** that is because Mr. Wren comes here every year!" Or **perhaps**, Sophie now feared, because he saw Langdon on television at the Vatican last year. (EO-FICTION-001)

(5) None of such marks may be used in connection with any other product or service in a manner that is **likely** to cause confusion among consumers, or to disparage or discredit the owner of such mark or its affiliates. (EO-WEBP-010)

(6) Brick hears machine guns, exploding grenades, and under it all, **no doubt** miles away, a dull chorus of howling human voices … (EO-FICTION-014)

Other meanings associated with this modality are the category of "doubt" [DO], and the category of "apprehension" [AP]. The former expresses uncertainty or lack of knowledge about the truth of the proposition, without assigning any degree of probability to it, as in (7):

(7) As you can see, we have introduced many policy measures which focus on quality rather than quantity. However, we have to ask ourselves whether we have done enough. I doubt it. (EO-SPEECH-002)

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

“Apprehension” is also known as “epistemic anxiety”, and is uncertainty combined with a positive or negative wish for or against the truth of a proposition, as in (8):

(8) If, at the end of it all, the reader remains unpersuaded by the less conventional of the arguments that I am trying to express, it is at least my hope that she or he will come away with something of genuine value from this tortuous but, I hope, fascinating journey. (EO-EXPE-004)

The tags proposed for deontic meanings are [OB] for obligation, [PRO] for prohibition, [RE] for recommendation, [PE] for permission, and [ABS] for absence of obligation. Under [OB], we include those meanings stemming from social laws and conventions of human interaction, as in (9):

(9) We **must** work for a result that means something for traders on the ground. (EO-SPEECH-003)

Under [PRO] we refer to cases of negative obligation in the sense of prohibition, as in (10):

(10) You **must** not collect or harvest any personal data of any user of Google Play or of any user of other Google Services via Google Play, including account names. (EO-WEBP-005)

The tag [RE] is used for cases of recommendation, that is, those states or events which are socially desirable, but not obligatory, because the deontic source is not entitled to or does not feel it necessary to impose their recommendation, as in (11):

(11) Perhaps many people who give $1,000 really **ought to** give at least $5,000, but to blame them for not giving more could be counterproductive. (EO-ESSAY-010)

For cases of permission, we propose using the tag [PE], as illustrated by (12):

(12) In these cases, to process your transaction and maintain your account, we **may** share your personal information with the product provider, as **permitted** under the Wallet Privacy Notice. (EOWEBP-005)

The tag [ABS] was used for those deontic meanings referring to absence of obligation, that is, indicating that there is neither obligation nor prohibition to perform a certain event, as in (13):

(13) One **needn’t** subscribe to any sort of irrationalist dogma to consent to the proposition “In the beginning was the Word.” (ETrans- ESSAY-002)

Dynamic modality concerns possibility and necessity derived from natural laws, as opposed to deontic modality, which is derived from social laws. For dynamic

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

modality, the general tag [DY] was used, with three main subtypes: "necessity" [NE], "tendency" [TE], and "possibility" [POS]. Some illustrative examples for each of these are provided in (14), (15), and (16), respectively:

(14) The calling to account of world leaders which this special Assembly represents is therefore not only most timely, but absolutely **necessary** for the sustainability and viability of our world. (ETrans-SPEECH-003)

(15) Because the world has a certain stability and doesn't change capriciously, the genes that have survived in the past **tend to** be the ones that are going to be good at surviving in the future. (EO-EXPE-006)

(16) In some districts half the children born **can** be expected to die before their fifth birthday. (EO-ESSAY-010)

The last tag in Figure 1 above represents volitional modality [VO] and refers to the expression of wish. There are two main subtypes under volitional modality: a) those meanings expressing "willingness" and "intention", which is captured with the tag [WI], as illustrated by (17) and (18), respectively:

(17) As I said a few days ago, we are **willing** to cooperate with the United States at the multilateral and bilateral levels (ETrans- SPEECH-001)

(18) I **am going to** do something unusual today. (EO-SPEECH-012)

(b) those meanings expressing the acceptance of both the occurrence and the non-occurrence of a state or event, which is captured with the tag [AC], as illustrated in (19):

(19) Then I don't **mind** if I am in the same room as them (EOFICTION-009)

In spite of the difficulties in distinguishing between these categories, the annotation experiments performed on two data sets (one containing two hundred English sentences and the other two hundred Spanish sentences) yielded a good degree of agreement between annotators (Kappa = 0.854), which indicates that the proposed tagsets for different modal meanings are reliable and consistent and can be used for large-scale annotation of the bilingual corpus. However, the annotation experiments also revealed difficult cases and problems in the annotation process. In some cases, there was a degree of overlap in the modal meanings expressed by the triggers, such as modal auxiliaries (*can, may, might, must*, etc.) in English and their counterparts in Spanish (*poder, deber, tener que*), as well as with some related adjectives (*possible*). These items are polysemous, that is, they tend to express more than one modal meaning (*must*: obligation, necessity, prohibition; *can*: permission, ability, situational possibility, prohibition), and this can give rise to potential

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

disagreement between annotators. One area of disagreement involving dynamic modality is illustrated by example (20):

(20) Port Aransas is the only established town on Mustang Island, and it's the fishing capital of Texas. Plenty of guides and charters are ready to take you out and demonstrate why; four lighted piers allow fishing by day and night. (EO-TOU-101)

In this example, the adjective "ready" is used to show willingness rather than mere ability.

In other cases, annotators disagreed on the modal nature of the triggers. This group mostly includes lexical verbs, adjectives, and nouns such as *prohibit*, *necessary,* or *obligation,* which have a meaning that is closely related to one of the modality types. Annotation experiments showed that the real challenge with these triggers is deciding whether they express a modal meaning or not, because these words are little or not grammaticalized at all, and while some of their uses are equivalent to modal constructions, as in (21), other uses are clearly non-modal, as in (22):

(21) We feel ourselves to be under a greater **obligation** to help those whose misfortunes we have caused. (EO-ESSAY-010)

(22) The United States is not living up to its **obligations** (EO-ESSAY-003)

The tool used for the annotation of modal meanings in English and Spanish was the UAM CorpusTool 3, given its user-friendliness and the additional functionalities it provides for quantitative analysis.[4]

6. Annotating metadiscourse markers in English and Spanish

The theoretical framework used to annotate metadiscourse markers was Hyland and Tse's characterization of these markers as a useful linguistic resource that writers use to communicate their stance and attitude towards a given proposition to their readers, thus emphasizing their interactional perspective (see Hyland & Tse 2004).

Interactive (textual) markers are concerned with ways of organizing discourse to anticipate readers' knowledge and include *transitions, frame markers, endophoric markers, evidentials,* and *code glosses.* Interactional (interpersonal) markers focus on the participants of the interaction and "seek to display the writer's persona and a tenor consistent with the norms of the disciplinary community" (Hyland & Tse

[4] The UAM Corpus Tool is freely available at <http://www.corpustool.com/>, and allows both automatic and manual annotation.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

2004: 139). These include *hedges, boosters, attitude markers, engagement markers,* and *self-mention markers.*

For interactional discourse markers (IDMs), an annotation scheme was designed consisting of a core tagset, including coarser annotation tags for IDMs, and and an extended tagset where finer-grained tags are specified. Table 6 graphically presents these tagsets:

Table 6. Core and extended tagset for IDMs in English and Spanish

Core tagset	**Extended tagset**
Stance Markers [SM]	Hedges [HE]
	Boosters [BO]
	Attitude [AT]
	Self-mention [SE]
	Other Stance [OS]
Engagement Markers [EM]	Questions [QU]
	Inclusive 1st person plural [1P]
	Indefinite 2nd person [2P]
	Directives [DI]
	Other Engagement [OE]

As in the case of modal meanings, the initial annotation scheme is hierarchical in order to allow annotators to choose more general or coarser tags when in doubt about the more specific ones. For example, if the coder is uncertain about which tag to assign to a markable s/he can simply tag it as "*Other Stance*" or *"Other Engagement"*. The abbreviated form of each tag is given in capital letters in brackets next to the full form.

As for the linguistic triggers, that is, the specific linguistic elements which realize or encode the different components of the proposed scheme, an initial taxonomy of potential linguistic realizations of IDMs in English and Spanish was designed, as specified in Table 7.

The list of potential triggers in Table 7 is purely illustrative, not exhaustive, and captures linguistic realizations extracted from the three genres of our bilingual corpus.

A pilot agreement study was performed on a training corpus of eighteen comparable texts extracted from the larger set of bilingual newspaper texts contained in the MULTINOT corpus. The training set contained equal proportions of the three newspaper genres (six news reports, six editorials, six letters to the editor), evenly divided into comparable data sets of texts written in English and in Spanish.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 7. Linguistic realizations of IDMs in English and Spanish (after Lavid & Moratón 2018)

LG	SF	English			Spanish		
		STANCE		ENGAGEMENT	STANCE		ENGAGEMENT
Episte-mic Adverb [EA]	Modal Adjunct	*Hedge: perhaps, possibly*	*Booster: Definitely, true*		*Hedge: quizás,*	*Booster: Definitivamente, la verdad*	
Attitude Adverb [ATA]	Disjunct	*Attitude: unfortunately, surprisingly*			*Attitude: desgraciadamente*		
[AJ]	Predicative Adjective in impersonal matrix clause	*Hedge: It is possible, likely + that*	*Booster: It is clear that*	---	*Hedge: Es posible que + Subj*	*Booster: Está claro que/es seguro que*	---
		Attitude: It is unfortunate that			*Attitude: Es lamentable que*		
	Predicative Adjective in interpersonal, speaker-hearer matrix clause	*Hedge John is unsure, doubtful whether*	*Booster John is sure, certain that*	---	*Hedge Juan está dudoso de que + Subjunctive*	*Booster Juan está seguro(a) de que + Subjunctive*	---
		Attitude: John is glad that			*Attitude: Juana está muy contenta de que …*		
	Predicative Adjective in to + infinitive clause	*Hedge: He is likely to have known*	*Booster: He is sure to have known*	---	----	-----	---
	Attributive Adjective in NG		*Booster: He is a sure winner*			*Booster: Es una apuesta segura*	
		Attitude: She is fantastic singer			*Attitude: Es una cantante fantástica*		

(*continued*)

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 7. (*continued*)

LG	SF	English		Spanish	
[N]	Noun complement in impersonal matrix clause	*HEDGE: There is a possibility/chance/likelyhood that + Indicative*		*HEDGE: Hay (una) posibilidad(es) de que + Subjunctive*	
[V]	Epistemic Verbal Operator in matrix clause [EPVO]	*HEDGE: It might be true It must be true*		*HEDGE: Podría ser verdad. Debe (de) ser verdad*	
	Deontic Verbal Operator in matrix clause [DEVO]		*He needs to act fast*		*Tiene que actuar rápido*
	Imperative Verbal form (Directive)		*DIRECTIVE: Tell that to my parents*		*Díselo a mis padres*
	Verbal inflection	----	---	*Future tense: e.g: Será verdad*	
[P]	1st P Pronoun	*SELF-MENTION I was a poll clerk*		*SELF-MENTION Yo me encuentro entristecido*	
	Incluse 1st P Pronoun as Subject		*We need to act*		*Tenemos que actuar*
	Indefinite 2nd P Pronoun as Subject		*You cannot seek*		*Ustedes …*
[CL]	Interrogative Clause as Rhetorical Question				

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

The annotations were carried out by two expert coders who tagged both data sets independently.[5]

The K value result obtained in the inter-annotator agreement study was 0.717 for the Spanish data set and 0.767 for the English data set, which can be interpreted as "substantial" agreement between coders. However, a thorough inspection of the annotations by both coders also shows areas of disagreement that merit investigation. In the majority of cases, disagreements occurred when one coder selected one option and the other one chose [none], which might be due to lack of attention and not necessarily from insufficient precision in the definition of the tags. The overall result is similar in both languages, in terms of both percentage and K value. The main conclusion is that, even though there is still room for improvement, the distinctions between the two main core tags (Stance and Engagement) and between their different subtypes are reasonably robust and can largely be replicated by different coders.

The overall results of this annotation study also revealed genre-related and language-specific variation in the distribution of IDMs in the three newspaper genres used for annotation, which can be considered a reflection of the different styles in newspaper writing in the British and Spanish communities.

7. Summary and concluding remarks

The annotation of discourse phenomena is important not only from a contrastive and translational point of view, but also for its computational relevance for the NLP community, where corpora enriched with annotations of complex linguistic features, such as thematization, modality, and metadiscourse markers, are an indispensable resource as input for a statistical learning approach to identify thematic, modal, and cohesive features of different genres, especially in applications concerned with the detection of document structure for automatic summarization or for the interpretation and generation of different text types. When the annotation of these features is bilingual or multilingual, as illustrated by the MULTINOT corpus, it has an added functionality and relevance for the improvement of current computational systems in different subfields, such as machine translation or automated natural language generation.

[5] See Lavid & Moratón (2018) for a detailed description of the annotation procedure.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Acknowledgement

Institute of Modern Languages and Translators (IULMYT).

References

Arús, Jorge, Lavid, Julia & Moratón, Lara. 2012. Annotating thematic features in English and Spanish: A contrastive corpus-based study. *Linguistics and the Human Sciences* 6: 173–192

Baker, Kathryn, Bloodgood, Michael, Dorr, Bonnie J., Callison-Burch, Chris, Filardo, Nathaniel W., Piatko, Christine, Lori & Miller, Scott. 2012. Use of modality and negation in semantically informed syntactic MT. *Computational Linguistics* 38: 1–28. https://doi.org/10.1162/COLI_a_00099

Boye, Kasper. 2012. *Epistemic Meaning: A Crosslinguistic and Functional-cognitive study* [Empirical Approaches to Language Typology 43]. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110219036

Correia, Rui, Mamede, Nuno, Baptista, Jorge & Eskenazi, Maxine. 2016. MetaTED: A corpus of metadiscourse for spoken language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis (eds), 3907–3913. <http://www.inesc-id.pt/publications/13153/pdf> (20 July 2017).

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Hendrickx, Iris, Mendes, Amália & Mencarelli, Silvia. 2012. Modality in text: A proposal for corpus annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC 2012*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 1805–1812. Istanbul: European Language Resources Association.

Hovy, Eduard & Lavid, Julia. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation* 22(1):13–36.

Hyland, Ken & Tse, Polly. 2004. Metadiscourse in academic writing: A reappraisal. *Applied Linguistics* 25(2): 156–177. https://doi.org/10.1093/applin/25.2.156

Krippendorff, Klaus. 2004. Reliability in Content Analysis: Some common Misconceptions and Recommendations. *Human Communication Research* 30(3): 411–433. <http://repository.upenn.edu/asc_papers/242> (12 Nomvember 2018).

Lavid, Julia, Arús, Jorge & Zamorano, Juan R. 2010. *Systemic-Functional Grammar of Spanish: a Contrastive Account with English*. London: Continuum.

Lavid, Julia & Moratón, Lara. 2015. Intersubjective positioning and thematisation in English and Spanish: A contrastive analysis of letters to the editor. *Nordic Journal of English Studies* 14 (1): 289–319.

Lavid, Julia & Moratón, Lara. 2016. Generic structures, rhetorical relations and thematic patterns in English and Spanish journalistic texts: A comparative study. (Paper presented at the 26th ESFLW).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Lavid, Julia & Moratón, Lara. 2018. Contrastive annotation of interactional discourse markers in English and Spanish newspaper texts. In *The Construction of Discourse as Verbal Interaction* [Pragmatics & Beyond New Series 296], Maria Ángeles Gómez González & J. Lachlan McKenzie (eds) 75–108. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.296.04lav

Lavid, Julia, Carretero, Marta, Arús Hita, Jorge, Moratón, Lara & Zamorano-Mansilla, Juan Rafael. 2014. Contrastive corpus annotation in the CONTRANOT Project: issues and problems. In *The Functional Perspective on Language and Discourse. Applications and Implications* [Pragmatics & Beyond New Series 296], Maria Ángeles Gómez González, Francisco José Ruiz de Mendoza Ibáñez, Francisco Gonzálvez-García & Angela Downing (eds), 57–86. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.247.04lav

Lavid, Julia & Moratón, Lara. 2016. Annotating metadiscourse markers in the English-Spanish MULTINOT corpus: Preliminary Steps. In *Conference Handbook of TextLink – Structuring Discourse in Multilingual Europe Second Action Conference*, Liesbeth Degand, Csilla Dér, Péter Furkó, Bonnie Webber (eds), 79–81. Debrecen: Debrecen University Press.

Lavid, Julia, Arús, Jorge & Moratón, Lara. 2013. Investigating thematic choices in two newspaper genres: An SFL-based analysis. In *Choice in Language: Applications in Text Analysis*, Gerard O' Grady & Lise Fontaine (eds), 187–214. London: Equinox.

Lavid, Julia, Carretero, Marta & Zamorano, Juan R. 2016a. Contrastive annotation of epistemicity in the multinot project: preliminary steps. In *Proceedings of the ISA-12, Twelfth Joint ACL – ISO Workshop on Interoperable Semantic Annotation, held in conjunction with Language Resources and Evaluation Conference 2016*, Harry Bunt (ed.), 81–88. <https://sigsem.uvt.nl/isa13/ISA-13_proceedings.pdf> (20 July 2017).

Lavid, Julia, Carretero, Marta & Zamorano Juan R. 2016b. A linguistically-motivated annotation model of modality in English and Spanish: Insights from MULTINOT. *Linguistic Issues in Language Technology* 14(4): 1–35. Standford CA: CSLI. <http://csli-lilt.stanford.edu/ojs/index.php/LiLT/article/view/67> (20 July 2017).

McShane, Marjorie, Nirenburg, Sergei & Zacharski, Ron. 2004. Mood and modality: out of theory and into the fray. *Natural Language Engineering* 10(1): 57–89.

Mora, Natalia. 2017. *Annotating Appraisal in English and Spanish Product Reviews from Mobile Application Stores: A Contrastive Study for Linguistic and Computational Purposes*. PhD dissertation, Universidad Complutense de Madrid.

Nissim, Malvina, Pietrandrea, Paola, Sansò, Andrea & Mauri, Caterina. 2013. Cross-linguistic annotation of modality: A data-driven hierarchical model. In *Proceedings of the 9th Joint ISO – ACL SIGSEM Workshop on Interoperable Semantic Annotation*, (isa-9) Harry Bunt (ed.), 7–14. Potsdam. <http://aclweb.org/anthology/W13-05> (20 July 2017).

Saurí, Roser & Pustejovsky, James. 2009. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation* 43(3): 227–268. https://doi.org/10.1007/s10579-009-9089-9

Szarvas, György, Vincze, Veronika, Farkas, Richárd & Csirik, János. 2008. The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, 38–45, Columbus OH: Association for Computational Linguistics. <http://www.aclweb.org/anthology/W08-0606> (20 July 2017).

Trnavac, Radoslava, Das, Debopam & Taboada, Maite. 2016. Discourse relations and evaluation. *Corpora* 11(2): 169–190. https://doi.org/10.3366/cor.2016.0091

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Taboada, Maite. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2: 325–347. https://doi.org/10.1146/annurev-linguistics-011415-040518

Van de Kauter, Marjan, Coorman, Geert, Lefever, Els, Desmet, Bart, Macken, Lieve & Hoste, Veronique. 2013. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal* 3: 103–120.

Wiebe, Janyce, Wilson, Theresa & Cardie, Claire. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 73(2–3): 165–210. https://doi.org/10.1007/s10579-005-7880-9

Zeyrek, Deniz, Demirşahin, Işın, Sevdik-Callı, Ayışığı & Cakıcı, Ruket. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse* 4(2): 174–184. https://doi.org/10.5087/dad.2013.208

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

PEST

A parallel electronic corpus of state treaties

Mikhail Mikhailov,[1] Miia Santalahti[2] and Julia Souma[2]
[1]University of Tampere, Finland / [2]University of Tampere

This chapter introduces the Parallel Electronic corpus of State Treaties (PEST). The current plan is to compile a parallel corpus, which will include treaties concluded between Russia and Finland, Finland and Sweden, and Sweden and Russia. In addition, there will be a subcorpus of international conventions in all three languages plus English, to be used as reference data. The chapter describes the structure of the subcorpora (number of documents, their chronological distribution and topics featured), and it also addresses the challenges of balancing such a corpus. In the future, this material can be used for studies ranging from lexicon and semantics to grammar, style, discourse, translation studies, and language for special purposes.

Keywords: legal language, language of state treaties, balanced corpus, compiling parallel corpora

1. General

Over the last decade, the number of text corpora in all imaginable languages has steadily increased, and many of these corpora are freely available online. Nowadays, the technical tasks related to corpora compilation (downloading texts, scanning, optical character recognition, aligning parallel texts, parsing, etc.) have become much easier. However, it seems that parallel and other kinds of multilingual corpora are still lagging behind. They remain much smaller than their monolingual counterparts and for many languages there are not any corpora at all. Moreover, the use of parallel corpora is often limited to parallel concordancing, while other query functions (frequency lists, collocations, N-grams, etc.) are much less popular.

To narrow the existing gap, a new corpus project was started at the University of Tampere in 2015. The objective is to collect a multilingual aligned parallel corpus

https://doi.org/10.1075/scl.90.11mik
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

of state treaties. International treaties form a parallel text material of (relatively) good quality. Most of these documents are publicly available, and there are no copyright issues related to them.

The state treaty is an important text: it is a contract document whose effect extends over the entire nation that is a party to it. Treaties have long served as one of several sources of international law, and nowadays are used increasingly as the dominant one (Hollis et al. 2005: 1). It can be said that treaties are no longer just contracts between governments, they are also instruments of law-making (Bunn-Livingstone 2002: vii).

As a linguistic entity, the bilateral treaty text is often of hybrid nature: during the treaty negotiation process, the text is written in both languages and translated back and forth so that the source and target language can no longer be identified (Probirskaja 2009: 61). Multilateral treaties (such as EU or UN treaties), on the other hand, are first drafted in certain languages (or one language) and later translated into other languages. This, naturally, gives rise to such research questions as whether differences can be found in the language of hybrid and translated treaties. However, treaty texts as objects of study are not favoured by linguists, but rather by researchers in the fields of political science, history and law, who are not particularly interested in the language of the texts. Non-linguists believe that different language versions of each treaty are identical. Fernando Prieto claims that quite commonly a working group chooses to work with the version whose language is more familiar to the users, without taking any notice of other language versions (Prieto Ramos 2011).

The Vienna Convention on the Law of Treaties (1969) stipulates that all language versions of a treaty should be explicitly pronounced authentic and equally authoritative: “When a treaty has been authenticated in two or more languages, the text is equally authoritative in each language, unless the treaty provides or the parties agree that, in case of divergence, a particular text shall prevail” (Article 33). The same document also suggests that “[t]he terms of the treaty are presumed to have the same meaning in each authentic text” (ibid.). However, from the linguistic point of view, nobody can guarantee that the meanings of all sentences in texts in different languages are identical. Therefore, it would be interesting to study what kinds of differences can be found in different language versions of international treaties and to what extent those differences can be deemed acceptable. This can be done by means of quantitative analyses of data from parallel corpora. Undoubtedly, the language of state treaties deserves attention from linguists and translatologists.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

2. PEST corpus

The corpus that is currently being compiled at the University of Tampere is entitled Parallel Electronic corpus of State Treaties (PEST). The project started in 2015. The current plan is to compile a parallel corpus, which will include treaties concluded between Russia and Finland (Russian <-> Finnish), Finland and Sweden (Finnish <-> Swedish), and Sweden and Russia (Russian <-> Swedish). In addition, there will also be a subcorpus of international conventions in all three languages plus English, and this subcorpus will be used as reference data. The estimated size of the whole corpus would be about 2 million running words. Thus, the corpus will have a recyprocal and harmonious structure (see Figure 1).

PEST is a full-text corpus, however, the annexes of the documents are not included (annexes usually consist of tables, figures and long lists that are not helpful for studying language, style, and translation issues). It is a diachronic corpus with data starting from 1917 (the year of the revolution in Russia and the independence of Finland) and extending to the present day. Expired treaties are also included, because the validity of treaties is not relevant from the point of view of linguistics, translation studies, and language technologies.

All versions of the documents are aligned at the sentence level and morphologically and syntactically annotated. Aligning and annotation are performed using open-source software (LF Aligner, TreeTagger, Finnish-dep-parser etc.). The corpus is stored on the server mustikka.uta.fi and a separate user account

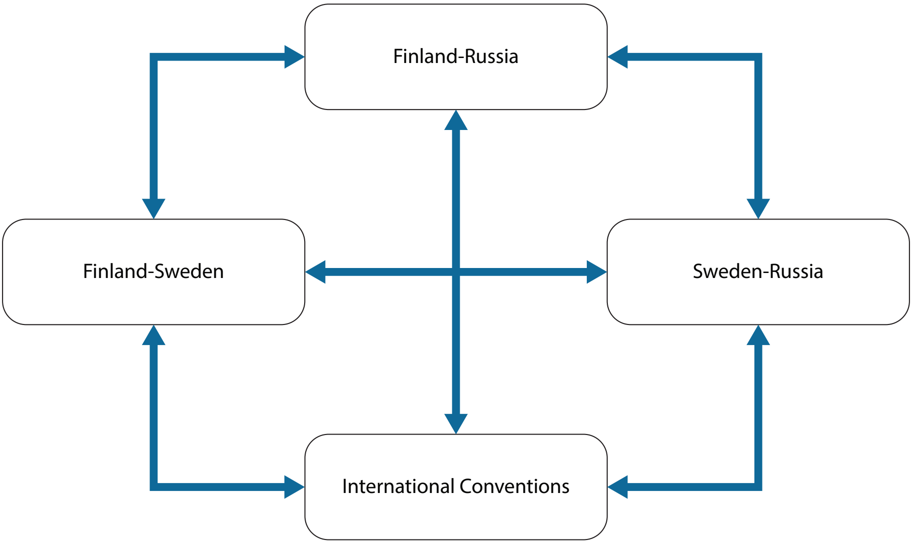


Figure 1. Structure of PEST

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

is required to access it. For the time being, the corpus is mainly intended for the use of the research group working at the University of Tampere, but access can be granted to others upon request. Once completed, the corpus may be opened for a larger-scale use within the research community. For corpus queries, the research group uses its own online corpus manager TextHammer <mustikka.uta.fi/texthammer>, which is specifically designed for working with parallel corpora. The software includes various tools, such as parallel concordances, frequency lists, N-grams, collocations, etc., and new functionalities are being developed.

The main idea of our corpus research is to compare documents by period, by country, and by topic. Research questions we aim to address by means of studying the PEST corpus include the following:

- How can the language of treaties be described in general? The treaty is a legal text and, therefore, the language should be highly standardized, precise and uniform, but is this always the case? What kinds of linguistic features are typical of the language of treaties?
- How has the state treaty as a linguistic genre evolved over 100 years?
- Are there differences in a particular language as used in treaties with different parties; for example, is the Finnish used in Swedish–Finnish treaties different from that used in Russian–Finnish treaties? If differences are discovered, what can this be attributed to? Can it reflect the power relations between the contracting states?
- How is the impact of translation reflected in the language of treaties? It is our assumption that, despite the hybrid nature of bilateral treaties, traces of translation activity are nevertheless present in them. After all, translation always plays an important role in the treaty negotiation and preparation process, which makes the state treaty a particularly interesting object of research for translatologists.

3. The Russian–Finnish section: Can a parallel corpus be balanced?

The Russian–Finnish section of PEST was completed in November 2016: it has 228 pairs of texts with around 300,000 running words of Russian text and 250,000 words of Finnish text. The documents have been divided into three chronological periods according to important political events: the first period starts from 1917, when Finland gained independence from Russia, the second one starts from 1944 after signing the Armistice Treaty (United Kingdom and USSR vs. Finland), and the third one starts from the collapse of the USSR in 1991. The consequences of each such major turn in the countries' relations are immediately reflected in the

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

treaty material: many new treaties are ratified in the following years, and their terms may be very different from those of earlier treaties. The composition of the Russian – Finnish section is presented in Table 1.

Table 1. The Russian–Finnish section of PEST

Subcorpus	Period	Number of text pairs	Word count, Russian	Word count, Finnish
A	1918–44	46	81,337	67,291
B	1945–91	128	141,764	114,982
C	1992–	54	81,482	65,334
Total:		228	304,583	247,607

Figure 2 further illustrates the composition of this corpus section.

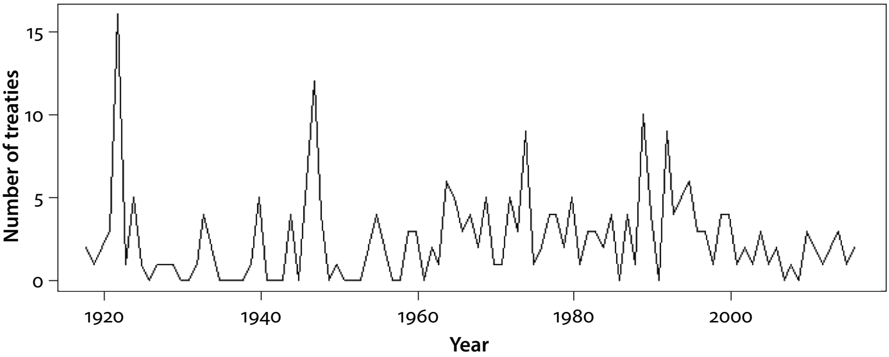


Figure 2. State treaties between Russia and Finland (1917–2016)

As we can see from the general statistics of this section of the corpus, there is an important methodological issue: the subcorpora are of different size, both in terms of the number of texts and word count, that is the length of the texts. Subcorpus C may still grow over time, as new treaties will belong to this subcorpus (until, perhaps, another radical change in relations takes place and subcorpus D becomes relevant), but the other two will not change in size. As a result, the classical principles of the proportionality and balancedness of a corpus (see for example. Biber 1993) are not adhered to.

It seems, however, that parallel corpora are a somewhat different type of data compared with monolingual corpora of general language (cf. Mikhailov 2017; Zanettin 2012). It is not easy to have all text types represented even in a monolingual corpus, but with multilingual data everything is multiplied: not all genres and topics exist in both languages and, what is more, not all kinds of texts are translated from one language into another. Therefore, a parallel corpus is much

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

more limited than a monolingual corpus. Let us remember that Noam Chomsky was critical in the question of representativeness of language corpora from the very beginning and his belief that all corpora are skewed (see for example, Andor 2004) cannot be completely ignored.

While it is possible to tackle the issue of skewedness if plenty of data is available, it is almost impossible to deal with it if the available data is limited, like in the case of our PEST project. To make the Russian–Finnish corpus chronologically balanced, we would have to remove texts from subcorpus B, which could lead to the omission of some important topics represented in other subcorpora and, consequently, hamper the comparability of data. Furthermore, some topics may be represented in certain periods only, for example peace treaties can be found in subcorpus A only. Moreover, Russia–Finland treaties are only a part of the PEST corpus, and attempts of harmonizing this section would likely create further problems.

The solution is to collect as many documents as possible, equip them with diverse metadata and deploy flexible tools to query the data by subcorpora. Our corpus is dynamic (Olohan 2004: 44), it can be updated with new texts, and its structure is flexible.

4. Other sections of the corpus

Work on other sections of the corpus has just started. Surprisingly, we discovered that the availability of documents in electronic format is limited, and finding different language versions is an even greater challenge. With regard to documents available online, our current data collecting strategy is to collect the language versions of documents separately first and then match them, align them and add to the corpus. The printed versions of state treaties are usually published as parallel texts, but when the treaty is published in more that two languages, the "national" versions may be aligned with the "international" (French or English) one and not with each other. For example, according to the Finnish Guideline for International Treaties, conventions that are to be published in Finnish, Swedish, French and English are implemented in such a manner that the Finnish edition is aligned with the English text and the Swedish edition is aligned with the French text (Valtiosopimusopas 2012: 115).

The following common topics can be found in bilateral treaties between each pair of countries:

1. Air traffic.
2. Avoidance of double taxation.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

3. Borders.
4. Communications (post, telephone, telegraph, television).
5. Culture, science and education.
6. Fishing.
7. Law and human rights.
8. Law of the sea.
9. Peace treaties.
10. Peaceful uses of atomic energy.
11. Trade.

4.1 The Sweden–Finland section

A preliminary list of treaties between Finland and Sweden has been compiled, and the total number of treaties is 211, which is surprisingly close to the number of Russia–Finland treaties. Figure 3 illustrates the volume and distribution of these treaties over time.

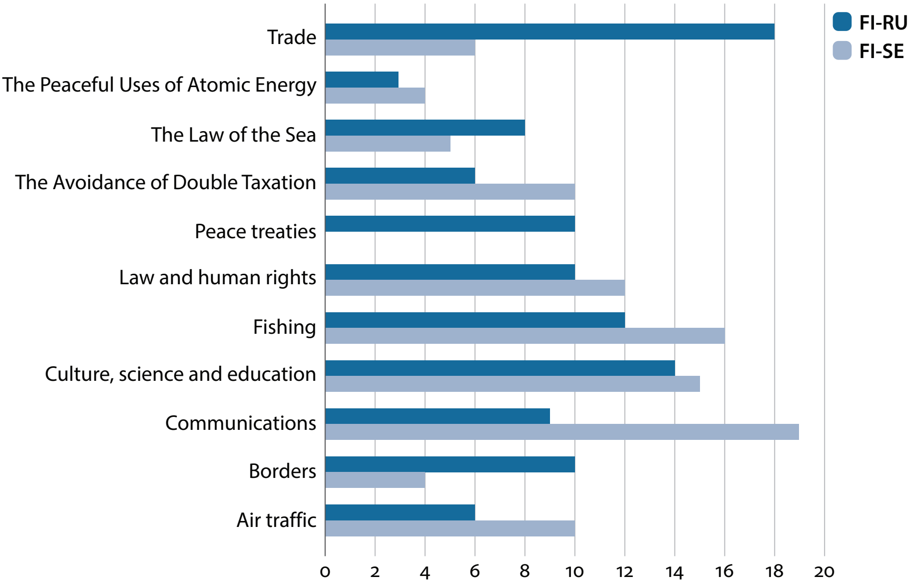


Figure 3. Treaties between Finland and Sweden (FI-SE) and Finland and Russia/Soviet Union (FI-RU) by topic.

Although the amount of treaties is almost equal, even a superficial study of the lists of treaties reveals a number of features specific of each pair of contracting states. For example, the following points stand out for Sweden–Finland treaties in particular:

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

- Many treaties between Sweden and Finland have been concluded by means of exchange of notes. Whether this is a feature typical of the Swedish treaty preparation process or a characteristic specific to the relations between Sweden and Finland in particular, we will have more information once the Sweden–Russia corpus is compiled.
- A number of treaties were concluded between the Grand Duchy of Finland (= part of the Russian Empire 1809–1917) and Sweden and later transformed into treaties between Finland and Sweden by exchange of notes.
- Many treaties are not bilateral but multilateral and include other Nordic countries (Sweden, Norway, Denmark and sometimes Iceland) (see Figure 4 below).
- Especially in the 1920s and 1930s, treaties were often concluded in three languages: Swedish, Finnish, and French, with a clause stating that the French version was to prevail in case of conflict between versions. There are a total of 15 such treaties, the earliest of which is from 1923[1] and the latest one from 1953 (see Figure 5).[2]
- A preliminary analysis conducted in conjunction with preparing the texts for adding to the corpus suggests that despite the almost equal total number of treaties, the size of the Sweden – Finland section in terms of running words is likely to remain smaller than that of the Russia–Finland section (exact data will only be available once all documents have been digitized). This is related to the fact that the Sweden–Finland treaties, in general, seem to be shorter. One explanation to this can be the large number of extended/transformed treaties (see above). Furthermore, the borders between Sweden and Finland have not changed significantly since 1917, which means that fewer detailed border treaties have been compiled between these countries than between Russia/Soviet Union and Finland, neither have Sweden and Finland been in war and thus they have not concluded any detailed peace treaties, either.[3]

[1] Suomen ja Ruotsin välinen selityskirja, tarkoittava työssä sattuvien tapaturmien korvausta [‘Agreement between Finland and Sweden regarding compensation for accidents occurring at work’].

[2] Suomen ja Ruotsin välinen sopimus lisäyksestä 29 päivänä tammikuuta 1926 tehtyyn sopimukseen riitaisuuksien sovinnollisesta ratkaisemisesta [‘Agreement between Finland and Sweden on an addition to the agreement on the conciliatory resolution of disputes from 29 January 1926’].

[3] The first treaty defining the borders between Finland and Sweden was the Treaty of Fredrikshamn of 1809. Since Finland became an independent state, the borders have been checked every 25 years and the treaties have been revised respectively. To date, such a revision has been performed four times: in 1926–1927, 1956–1957, 1981, and 2006.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

- As can be seen by comparing Figure 2 and Figure 3, the annual distribution of treaties is much more uniform between Sweden and Finland than between Russia/Soviet Union and Finland. The busiest year in terms of concluding treaties between Soviet Union and Finland was 1922, when a total of 16 treaties were signed, whereas the highest number of treaties concluded between Sweden and Finland in one year was 7 in 1933.
- Swedish is one of the two official languages of Finland, and therefore treaties between Finland and Sweden can be concluded in Swedish. Probably for this reason, a large number of Finland – Sweden treaties – not only diplomatic notes but regular treaties as well – are labelled as translations.[4] (The Finnish

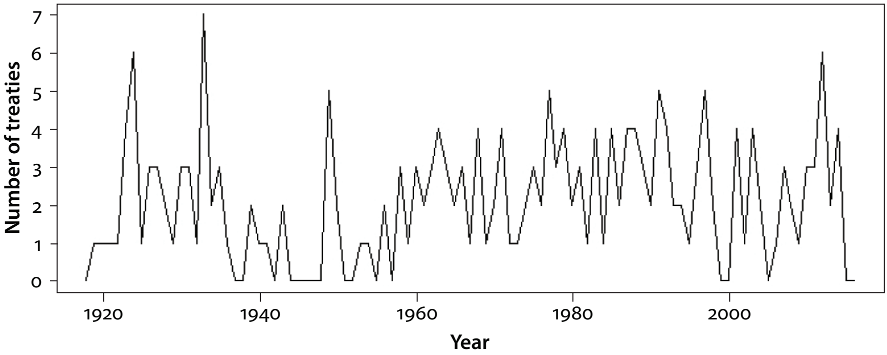


Figure 4. State treaties between Finland and Sweden (1917–2016)

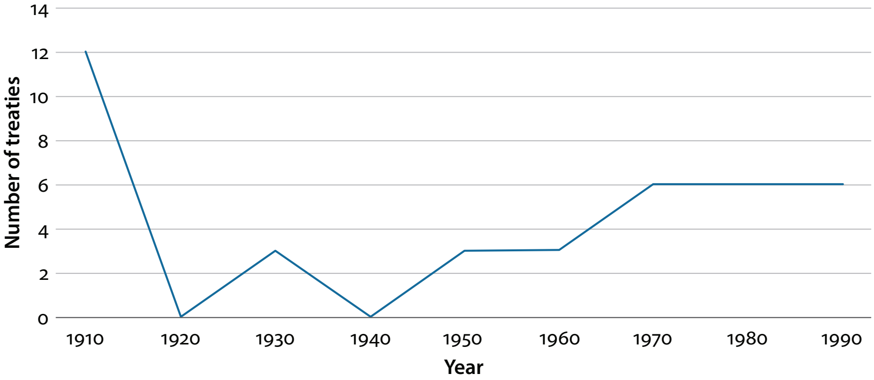


Figure 5. Multilateral treaties between Nordic countries (Finland, Sweden, Norway, and Denmark) 1918–2016

[4] The published version of the treaty has an explicit comment Suomennos (= translation into Finnish).

versions of Finland – Russia treaties are almost always proclaimed authentic, with the exception of exchange of notes and a few other documents).[5]

4.2 The Russia–Sweden section

The compilation of the Russia–Sweden section of the corpus has proven to be a surprisingly difficult task. Only a few treaties are available on the online service of the Russian Ministry of Foreign Affairs, and the series "Dokumenty vnešnej politiki SSSR" (Documents of Soviet Foreign Policy) only contains documents concluded before 31 December 1943. At the time of writing this chapter, a total of 41 treaties have been obtained for the Russia–Sweden section of the corpus. This section is expected to remain smaller than the Russia–Finland and Sweden–Finland sections for various reasons. For example, Sweden and Russia do not share a border, which immediately excludes the need for treaties on borderlines, border crossing points, activities in border zones, cross-border traffic, etc. For the sake of comparison, let it be noted that the number of treaties between Russia and Finland falling into these categories is 61. It can be expected that once more treaties between Russia and Sweden are obtained, texts dealing with communications (telephone, telegraph, mail, radio, television), the sea (the Baltic Sea in particular), and other such topics will also be represented.

4.3 Multilateral international treaties

The section of international treaties is different from the previous three. It will be used mainly as a reference corpus to discover features that are specific of bilateral treaties. Still, this corpus can also be used for stand-alone research. This is an exceptional type of material: there are documents available in great quantities, and this will be the only section of PEST, which features limited selection. Furthermore, this section will be multilingual, while the other sections are (mainly) bilingual. Multilingual alignment presents certain technical challenges but, fortunately, LF Aligner,[6] which is being used for the purpose, supports aligning more than two texts.

The collection of texts for this section may also prove to be more challenging than for the bilateral sections. Multilateral treaties conducted since the establishment of the UN in 1945 can be obtained from the UN's website, and treaties conducted within the Council of Europe are also available, but older documents

[5] These are cases where one of the parties initiates a treaty and suggests its text and, consequently, the other party produces a translation.

[6] <https://sourceforge.net/projects/aligner/>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

may be more difficult to find. One possible source is the League of Nations, which was in effect from 1919 to 1946. However, it must be taken into account that different countries have joined the above-mentioned organizations at different times; for example, Sweden was one of the founders of the League of Nations, Finland joined the League in December 1920, and the Soviet Union became a member in September 1934 but was discharged as soon as in December 1939. As for the UN, the Soviet Union was one of its founders and Russian is a working language of the organization, while Sweden became a member in 1946 and Finland in 1955. Therefore, it may be difficult to find early treaties in all four languages. Further possible sources of material include the Geneva Conventions, and treaties concluded by the International Red Cross and Red Crescent as well as other international organizations such as World Trade Organization, the Organization for the Prohibition of Chemical Weapons, International Atomic Energy Agency, etc.

5. Great expectations

PEST is a dynamic corpus, which means that its structure is flexible and it will change, as new data will be added. In other words, the corpus will never be exhaustively finalized, although the number of documents is not very large and it cannot be expanded indefinitely.

Even the Russia–Finland section, which is completed, does not contain all available treaty data. The current version is bilingual with Russian and Finnish. However, Finland has two official languages, and all treaties are also available in Swedish. The Swedish of Finland is different from the Swedish of Sweden, and thus the Russia – Finland treaties in Swedish could provide very interesting research material for comparisons with other sections of PEST, bearing in mind that these documents are translations into Swedish and not parallel language versions of the treaties. Moreover, some treaties between Russia and Finland also have French or English versions.

The research conducted with the corpus data will also be "dynamic", that is we shall not wait until the last text is added to the corpus. Instead, studies will be conducted all the time, and case studies performed on limited data will be retested on more extensive data or on other sections, if needed. The first pilot studies conducted using the corpus have already provided interesting findings. For example, a study of lemmatized Russian and Finnish frequency lists (Mikhailov & Santalahti 2017) suggests that the treaties between Russia/Soviet Union and Finland contain elements that can be considered untypical of the legal genre and even deemed as emotionally charged. Another small-scale pilot study (with a publication probably to come in 2018) revealed that language typical of the Soviet discourse, that is

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

ideologically oriented elements, could easily be found in treaties of subcorpus B. An interesting point discovered in both of the above-mentioned studies was that the Finnish language in the "emotionally" or "ideologically" charged segments was quite unidiomatic. This suggests that these segments were probably translated from Russian into Finnish and provides grounds for further studies in order to analyze the background of this phenomenon: can it be interpreted as a sign of the Russian party's dominance in the treaty negotiation process, is it a form of compromise, or just a mutually accepted feature in the treaty genre of that time, for example? Another study with the Russian–Finnish data from PEST material deals with negation in the texts of treaties (Souma et al. 2017). The scope of future studies on this material will range from lexicon and semantics to grammar, style, discourse and language for special purposes.

Acknowledgement

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources and SIU – Norwegian Centre for International Cooperation in Education for support (project NCM-RU-2016/10039).

References

Andor, József. 2004. The master and his performance: An interview with Noam Chomsky. *International Pragmatics* 1(1): 93–111.

Baker, Mona. 1996. *Corpus-based translation studies: The challenges that lie ahead*. In *Terminology, LSP and Translation*, Harold Somers (ed.), 175–187. John Benjamins. https://doi.org/10.1075/btl.18.17bak

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243–257. https://doi.org/10.1093/llc/8.4.243

Bunn-Livingstone, Sandra L. 2002. *Juricultural Pluralism Vis-À-Vis Treaty Law: State Practice and Attitudes*. Leiden: Brill.

Hollis, Duncan B. et al. 2005. *National Treaty Law and Practice*. Leiden: Brill.

Mikhailov, Mikhail. 2017. Are classical principles of corpus compiling applicable to parallel corpora of literary texts? In *Translation Studies and Translation Practice: Proceedings of the 2nd International TRANSLATA Conference, 2014* Part 1 [Forum Translationswissenschaft 19], Lew N. Zybatow, Andy Stauder & Michael Ustaszewski (eds), 151–157. Frankfurt: Peter Lang,

Mikhailov, Mikhail & Santalahti, Miia. 2017. Mistä kertoo valtiosopimusten kieli? Tapaustutkimus interferenssistä Suomen ja Venäjän välisissä valtiosopimuksissa (What can we tell from the language of state treaties? Case study on interference in treaties between Finland and Russia). *MikaEL, Electronic Journal of the KäTu Symposium on Translation and Interpreting Studies* 10: 73–87.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London: Routledge. https://doi.org/10.4324/9780203640005

Prieto Ramos, Fernando. 2011. El traductor como redactor de instrumentos jurídicos: el caso de los tratados internacionales (Translator as an editor of legal instruments: The case of international treaties). *JoSTrans* 15: 200–214.

Probirskaja, Svetlana. 2009. *Rajankäyntiä: Suomen ja Venäjän kahdenväliset valtiosopimukset käännöstieteellisen avainsana-analyysin valossa* (Across the Border: Finnish and Russian State Treaties in the Light of Keyword Analysis). PhD dissertation, Tampere University Press.

Souma, Julia, Kudashev, Igor & Mikhailov, Mikhail. 2017. Otricanie v russkih i finskih versiâh dvustoronnih dogovorov meždu Rossiej i Finlândiej: opyt korpusnogo issledovaniâ (Negation in Russian and Finnish versions of the state treaties between Russia and Finland: A corpus-based research). In *Conference Proceedings Computational Linguistics and Intellectual Technologies (Komp'juternaja lingvistika i intellektual'nye tehnologii). 23nd International Conference on Computational Linguistics and Intellectual Technologies – 2017*. Moscow.

Toury, Gideon. 2012. *Descriptive Translation Studies – and Beyond*, rev. edn. [Benjamins Translation Library 100]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.100

Valtiosopimusopas 2012. Kansainvälisten ja EU-sopimusten valmistelua ja voimaansaattamista koskevat ohjeet (Manual on preparation and ratification of international and EU treaties). Helsinki: Publications of the Finnish Ministry of Foreign Affairs.

Zanettin, Federico. 2012. *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Indexation and analysis of a parallel corpus using CQPweb

The COVALT PAR_ES Corpus (EN/FR/DE > ES)

Teresa Molés-Cases and Ulrike Oster
Universitat Politècnica de València, Universitat Jaume I

This contribution presents a section of the Corpus Valencià de Literatura Traduïda (COVALT), created by the research group of the same name (Department of Translation and Communication, Universitat Jaume I, Spain). The COVALT corpus is a four-million word corpus made up of narrative works originally written in English, French, and German and their Catalan translations published in the autonomous community of Valencia between 1990 and 2000. Since the members of the Covalt group are interested in translation research, and more specifically in the investigation of translated Catalan and Spanish, this corpus has recently been extended to include translations into Spanish published in Spain (COVALT PAR_ES corpus). This chapter presents the COVALT PAR_ES corpus, as well as its process of compilation and analysis with CQPweb.

Keywords: corpus compilation, corpus indexation, CQPweb, COVALT corpus

1. Introduction

The description of translated language fundamentally relies on two different types of corpora: comparable and parallel. There are various parallel corpora (some of which are combined with comparable corpora) available for different language pairs, many of which include English. Those based on literary texts and available to researchers include COMPARA (Frankenberg-Garcia & Santos 2003), ENPC (Oksefjell 1999), OMC (Johansson 2004), PaGeS (Doval et al., this volume), PELCRA (Przepiórkowski et al. 2010), ParFin, and ParRus (Mikhailov & Cooper 2016). Over the past decade, considerable efforts have been made to create resources to facilitate translation research related to Spain's co-official languages, Spanish, Catalan, Basque and Galician: see, for example, P-ACTRES (Sanjurjo-González & Izquierdo, this volume), CLUVI (Gómez Guinovart & Sacau Fontenla

https://doi.org/10.1075/scl.90.12mol
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

2004), Aleuska (Zubillaga, Sanz & Uribarri 2015) or IULA Spanish–English Technical Corpus (Badia et al. 1998). However, both the number and size of available resources for specific individual language combinations are still limited. While comparable corpora are relatively easy to compile and render accessible to interrogation through corpus management software, creating a freely accessible parallel corpus requires a degree of expertise and computer literacy that is not always easy for translation scholars to attain. IMS Open Corpus WorkBench, with its Corpus Query Processor (CQP) (Hardie 2014), is one of the most widely used tools for handling corpora (including parallel ones). The web-based graphic user interface CQPweb (Hardie 2012), which provides tools that have greatly enhanced the accessibility, flexibility and user-friendliness of corpora, was upgraded to a version capable of handling parallel corpora in mid-2016.[1] The present chapter provides a detailed overview of the creation of a section of the COVALT corpus: a parallel corpus of original narrative works in English, French and German, together with their translations into Spanish. We will first describe the characteristics of this subcorpus within the larger architecture of COVALT (Section 2), then provide a step-by-step description of the compilation and indexing process for CQPweb (Section 4) and finally suggest some ways of analysing it (Section 5).

2. The corpora

The COVALT corpus (Corpus Valencià de Literatura Traduïda) was created by the Covalt research group at Universitat Jaume I (Spain). It was originally created as a four-million-word multilingual parallel corpus containing complete narrative works originally written in English, French and German with their Catalan translations published in the autonomous region of Valencia between 1990 and 2000 (Guzman 2013: 49). The original COVALT corpus has since been extended to include three additional sections (cf. Figure 1):

- COMP_CAT: A comparable Catalan corpus of original narrative works published during the same period and in the same geographical region (Comunidad Valenciana).
- PAR_ES: A parallel corpus consisting of translations into peninsular Spanish of the same novels originally used for the EN/FR/DE > CAT corpus.
- COMP_ES: A comparable Spanish corpus of original narrative works published in Spain during approximately the same period.

[1] CQPweb was originally based on the online query interface of the well-known British National Corpus (BNCweb) (Hardie 2012).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

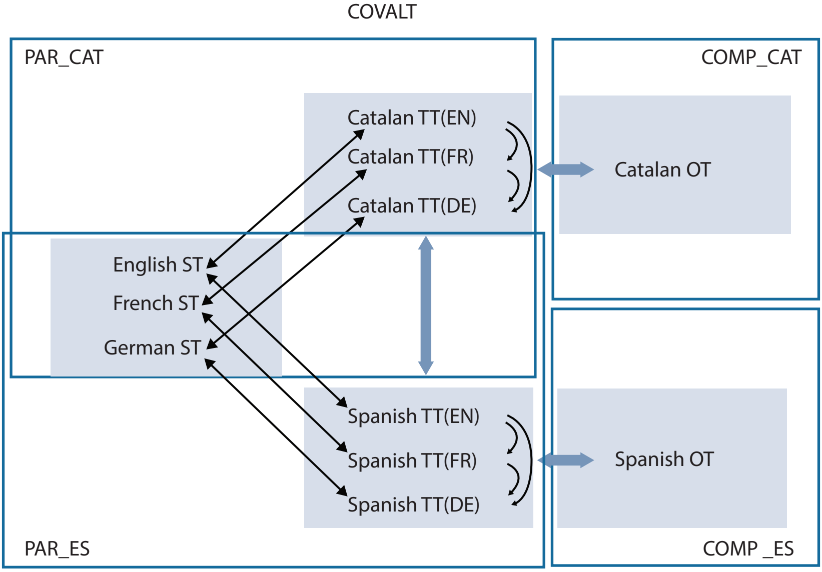


Figure 1. Architecture of the COVALT corpus

The four sections (PAR_CAT, PAR_ES, COMP_CAT AND COMP_ES), have been kept to approximately the same size (around 4 million words) to enhance comparability. This number was determined by the availability of works in the smallest of the defined populations: Catalan translations published in the autonomous region of Valencia between 1990 and 2000. During that period, more Catalan translations were published from English than from French and German and the sizes of the EN/FR/DE subcorpora reflect this imbalance. Future enlargements of the corpus will aim to compensate for this over time, while maintaining an overall balance in size between the sections.

The COVALT corpus[2] has expanded in response to the research interests of the members of the Covalt group, who have a special interest in translated Catalan and Spanish. It now allows for diverse approaches, including:

- Analyses of translated Catalan
- comparing Catalan target texts (CAT TT) with Catalan original texts (CAT OT): for example Marco (2013a, 2018b);
- comparing Catalan target texts with their source text corpora (English ST/French ST/German ST) (e.g. Guzman 2015a, 2015b, 2016; Marco 2018a, 2018b, 2013b; Oster and van Lawick 2013; Verdegal 2013, 2014);

[2] Access for research purposes can be requested via the Covalt group's webpage: <http://www.covalt.uji.es>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

- comparing Catalan target texts from different source languages with each other.
- Analyses of translated Spanish
- comparing Spanish target texts (ES TT) with Spanish original texts (ES OT) (Martínez Vilinsky 2016);
- comparing Spanish target texts with their source text corpora (DE ST) (Molés-Cases 2016a);
- comparing Spanish target texts from different source languages with each other.
- Comparison of Spanish and Catalan translated language (CAT TT vs. ES TT) (Oster & Molés-Cases 2016).

In this chapter, we will describe the newly created parallel corpus EN/FR/DE>ES, in order to elucidate the compilation and indexation process for CQPweb. That section of COVALT contains a total of 4,357,846 words divided into the following three subcorpora:

- The English>Spanish subcorpus includes 36 novels originally written in English and their translations into Spanish: a total of 2,417,463 words.
- The French>Spanish subcorpus consists of 21 novels: a total of 1,106,222 words.
- The German>Spanish subcorpus contains 14 novels: a total of 834,161 words.

3. Corpus compilation and indexation

The different sections of the COVALT corpus have been compiled and indexed in IMS Open Corpus WorkBench, a corpus analysis package developed by the Institut für Maschinelle Sprachverarbeitung in Stuttgart. The Covalt group chose this collection of open-source tools because it includes an advanced tool for corpus analysis, the Corpus Query Processor (CQP), which allows researchers to quickly perform complex searches of very large corpora by means of a sophisticated query language. Furthermore, since 2016, the online version of CQP (CQPweb)[3] has been able to handle both monolingual and parallel corpora indexed in IMS Open Corpus WorkBench (version 3.2.0 or later). The indexing process facilitated by this new tool (Hardie 2016) is slightly different from the traditional indexing process in IMS Open Corpus WorkBench (Evert et al. 2016, Molés-Cases 2016b). The Covalt group has therefore adapted the compilation and indexation of its corpora

[3] CQPweb is a web-based graphic user interface for some elements of the CWB, such as Corpus Query Processor.

to make them available for online analysis. In view of the complexity and novelty of the process, this section will provide a step-by-step explanation of the corpus compilation and, in particular, the indexing and encoding process for CQPweb.[4]

3.1 Preparation of texts

First, the parallel corpus needs to be prepared in compliance with the requirements of CQPweb, that is "laid out in plain-text files, in vertical format, with one token per line" (Hardie 2012: 390). In our case, this was done in five steps, employing specific tools at each stage (see Martínez Vilinsky 2016 and Molés-Cases 2016b for more detailed information on the compilation process):

- Scanning (OCR)
- Alignment of source and target fragments (Déjà Vu)[5]
- Splitting the exported translation memory (Perl script *separab3.pl*). This step was necessary because the exported translation memory is a bilingual text and taggers can only tag monolingual texts.
- Grammatical and lexical tagging (TreeTagger)[6]
- Metatextual tagging (Notepad++). These tags include bibliographical information such as title, author, publication date, publisher, etc.

The tagging process results in a verticalized text that can serve as input for the indexing process. Figure 2 shows the verticalized text file of the sentence "Folklore, legends, and fairy tales have followed childhood through the ages […]" from *The Wonderful Wizard of Oz*.

The first column includes each token of the corpus, followed by its corresponding grammatical category as indicated by specific tags (second column) and the lemma (third column).[7] In addition, XML tags for the beginning and end of sentence (<s > </s>) and text (<text> </text) appear in separate lines. In

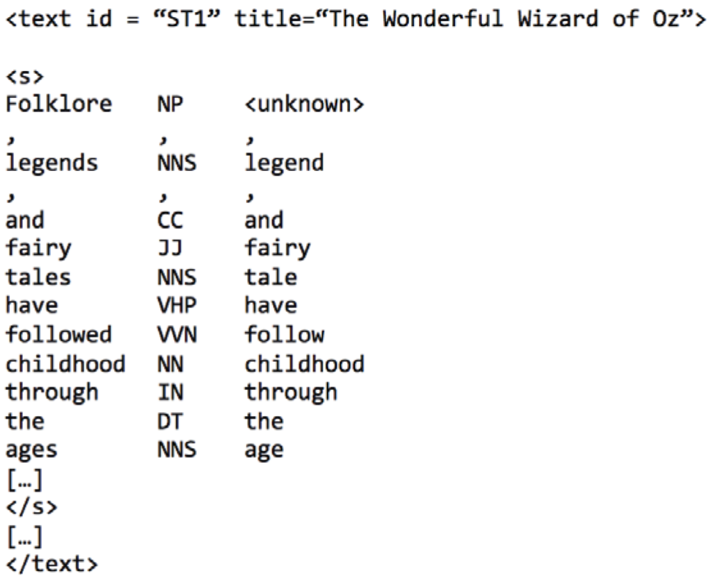
[4] The specific version used for this project was CQPweb v. 3.2.23. Download and installation information on CQPweb, CWB and CWB-Perl interface and the other necessary packages is available at Hardie (2016: 8–9) and <http://cwb.sourceforge.net/download.php#gui>.

[5] DéjàVu is only compatible with Windows. Other similar software applications include LF Aligner, Youalign, Stingray, Trados, Transit, etc.

[6] As tagging is language-specific, some programs work better for certain languages than others. TreeTagger was used here, while Freeling was the preferred option for Catalan.

[7] We can see an example of TreeTagger's margin of error in Figure 2, in which the word *folklore* has been incorrectly tagged as proper noun (NP) and lemmatized as an unknown lemma (<unknown>).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



```
<text id = "ST1" title="The Wonderful Wizard of Oz">

<s>
Folklore    NP     <unknown>
,           ,      ,
legends     NNS    legend
,           ,      ,
and         CC     and
fairy       JJ     fairy
tales       NNS    tale
have        VHP    have
followed    VVN    follow
childhood   NN     childhood
through     IN     through
the         DT     the
ages        NNS    age
[…]
</s>
[…]
</text>
```

Figure 2. Verticalized text

this version of CQPweb, the tags for beginning and end of corpus (<corpus> </corpus>) are unnecessary.

The main requirements for the tagged texts to be indexed are as follows:

- Text codification must be UTF8 without BOM. If necessary, the recodification can be done using the software application Notepad++.
- It is helpful to give all files the same, recognizable file extension. We used.*vert* (for "verticalized text"). This can be modified using Notepad++.
- The tagged text must not contain any hidden characters indicating carriage return (\r). These can be eliminated through a regular expression with Notepad++.
- TreeTagger introduces the tag "<unknown> "whenever it does not recognize a word (and therefore cannot establish its lemma). These tags must be removed, because <unknown> is not an established tag for CQP. We have replaced them with "unknown" (i.e. treated them as lemmas labelled "unknown", not as tags).

3.2 Uploading the files to CQPweb

The following subsections detail the steps required to complete the compilation process for a parallel corpus for analysis in CQPweb.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Step 1: Creating directories

First, the user has to create several directories on the server on which CQPweb has been installed (accessed via Putty[8] or command line).[9] In our case, these directories are:

(a) A folder for the corpora to be indexed (*indexing*):
```
    mk dir /home/user_name/indexing
```
(b) A folder for the indexed corpora (*corpora*):
```
    mk dir /home/user_name/corpora
```
(c) A folder for each subcorpus inside the previous folder. In this example, we use the following English–Spanish pair of corpora:

Corpus	Name of folder
English ST (original English texts)	ANGESPOR
Spanish TT (Spanish translated texts)	ANGESPTRAD

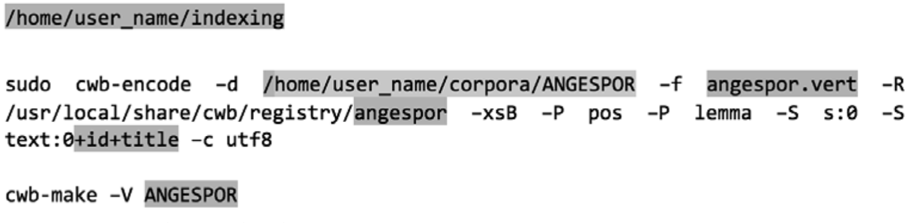
This results in the following command lines:

```
mk dir /home/user_name/corpora/ANGESPOR
mk dir /home/user_name/corpora/ANGESPTRAD
```

Step 2: Encoding and indexing corpora in CWB

The next step is to encode and index the corpus. Encoding a corpus in CWB involves "convert[ing] the verticalized text to CWB binary format with the cwb-tool" (Evert et al. 2016: 3). Indexing involves building the necessary index files.

Figure 3 shows the instructions that have to be typed in order to encode and index a corpus, with variable segments highlighted in grey. Each of the commands is followed by a line feed. We will take the ANGESPOR corpus as an example.



```
/home/user_name/indexing


sudo  cwb-encode  -d  /home/user_name/corpora/ANGESPOR  -f  angespor.vert  -R
/usr/local/share/cwb/registry/angespor  -xsB  -P  pos  -P  lemma  -S  s:0  -S
text:0+id+title -c utf8

cwb-make -V ANGESPOR
```

Figure 3. Encoding and indexing instructions

[8] Putty is an SSH and telnet client: <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>.

[9] A machine or server with a Unix-style operating system is needed to set up CQPweb.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 1 explains each function within the previous commands.

Table 1. Explanation of encoding and indexing instructions

Function	Explanation	Our case
sudo	administration permission	
cwb-encode	encoding instruction	
-d	directory where the corpus will be located	`/home/user_name/corpora/ANGESPOR`
-f	the file to be indexed	angespor.vert
-R	directory where the registry of the corpus will be written	`/usr/local/share/cwb/registry/angespor`*
-xsB	the corpus is organized in XML	
-P	corpus attributes	pos, lemma
-S	structural attributes	s: 0** text: 0 + id+title***
-c	codification	utf8
cwb-make	indexing instruction	

* *Note* that *angespor* is lower case here because it is a registry file and not the name of the corpus (ANGESPOR).

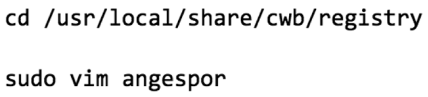
** The *s* refers to the sentence level; the *0* indicates that it is not possible to include one sentence within another sentence.

*** This means that the following metadata have been used for the COVALT corpora in CQPweb: text identification (ST1, TT1, etc.) and the title of each novel (*The Wizard of Oz, El mago de Oz*, etc.).

The same process has to be followed in order to encode and index the target subcorpus (ANGESPTRAD).

Step 3: Aligning the subcorpora

After having encoded and indexed the two components of a parallel corpus (in our example: ANGESPOR and ANGESPTRAD), the next step is to align these two subcorpora. Figure 4 outlines the instructions used to access the registry files in order to execute the text editing program that will allow for the alignment of the subcorpora.



```
cd /usr/local/share/cwb/registry

sudo vim angespor
```

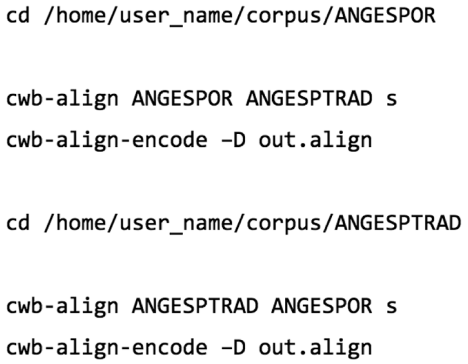
Figure 4. Executing text editing program files

Next, use the down arrow on the keyboard to scroll down to the end of the document and activate the editing option by typing "i". Type the name of the subcorpus

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

to be aligned: "ALIGNED angesptrad". Then press the Esc key and type ":wq" (*write and quit*).

Finally, the subcorpora ANGESPOR and ANGESPTRAD are aligned using the instructions *cwb-align* and *cwb-align-encode* (Figure 5), that is each sentence in the source subcorpus is linked to a corresponding sentence in the target subcorpus.



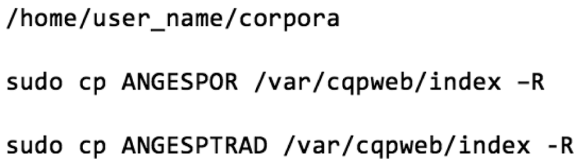
```
cd /home/user_name/corpus/ANGESPOR

cwb-align ANGESPOR ANGESPTRAD s
cwb-align-encode –D out.align

cd /home/user_name/corpus/ANGESPTRAD

cwb-align ANGESPTRAD ANGESPOR s
cwb-align-encode –D out.align
```

Figure 5. Alignment instructions

The "s" at the end of the command refers to "sentence" since the subcorpora are aligned on a sentence level (see above). The function – D places the generated files in the same directory as the corpus's other attributes (specified in the registry file) (Evert et al. 2016: 18).

Step 4: Copying the files to CQPweb

By this stage, both subcorpora have already been encoded and indexed. However, to be analysable online, the folders of each subcorpus must be copied (*cp*) to the CQPweb folder (cf. Figure 6).



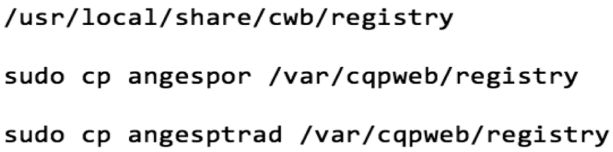
```
/home/user_name/corpora

sudo cp ANGESPOR /var/cqpweb/index –R

sudo cp ANGESPTRAD /var/cqpweb/index -R
```

Figure 6. Copying corpus folders to the CQPweb folder[10]

The same action must then be executed for the registry files (Figure 7).

[10] –*R* refers to the registry folder.

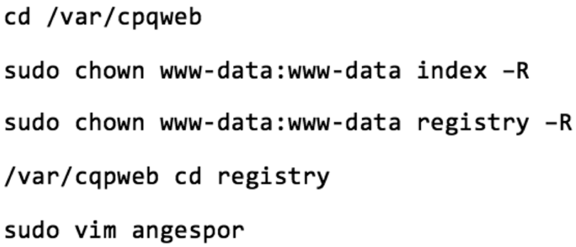
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



```
/usr/local/share/cwb/registry

sudo cp angespor /var/cqpweb/registry

sudo cp angesptrad /var/cqpweb/registry
```

Figure 7. Copying registry files

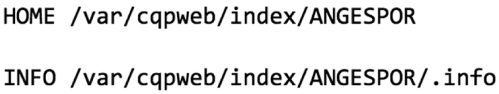
Once the previous steps are completed, the file permissions for each corpus must be copied. Figure 8 demonstrates this process for the source subcorpus ANGESPOR. The process must then be repeated for the target subcorpus.



```
cd /var/cpqweb

sudo chown www-data:www-data index -R

sudo chown www-data:www-data registry -R

/var/cqpweb cd registry

sudo vim angespor
```

Figure 8. Copying file permissions

Finally, once the corpus has been installed on CQPweb, the program must be told where *HOME* and *INFO* are located. Use the arrow on the keyboard to scroll down to *HOME* and *INFO* and replace the route with the one corresponding to CQPweb, as shown in Figure 9.



```
HOME /var/cqpweb/index/ANGESPOR

INFO /var/cqpweb/index/ANGESPOR/.info
```

Figure 9. Replacing route information for HOME and INFO

To exit, press the Esc key again and type ":wq" (*write and quit*).

Step 5: Activating the corpora on the web interface

Finally, the indexed corpora must be installed on CQPweb. First the user has to access CQPweb using their username and password, select the menu CQPweb's *Admin Control Panel* and activate the option *Install new corpus* (in the left margin of the *Admin Control Panel*) from the *Corpora* submenu (Figure 10).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

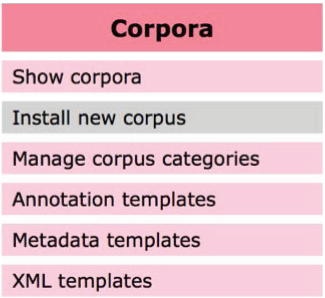
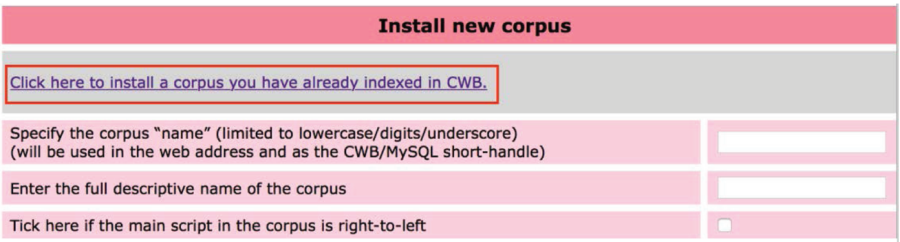


Figure 10. *CQPweb Admin Control Panel (Corpora)*

Since we have already indexed our corpus in CWB, we will select the option *Click here to install a corpus you have already indexed in CWB,*[11] as shown in Figure 11.



Install new corpus

Click here to install a corpus you have already indexed in CWB.

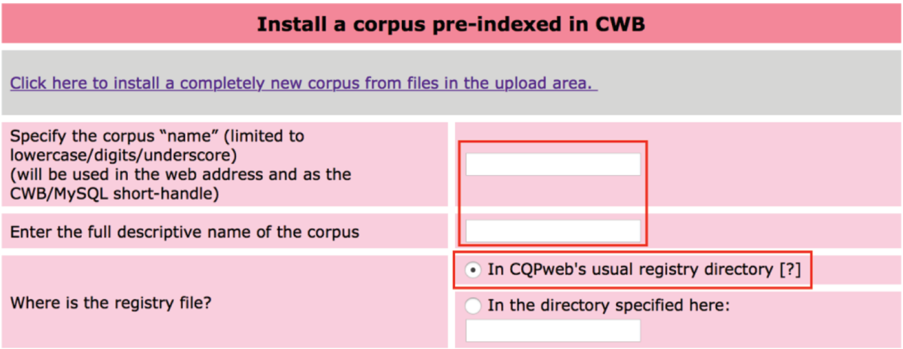
Specify the corpus "name" (limited to lowercase/digits/underscore)
(will be used in the web address and as the CWB/MySQL short-handle)

Enter the full descriptive name of the corpus

Tick here if the main script in the corpus is right-to-left □

Figure 11. *CQPweb Admin Control Panel (install new corpus)*

We then need to specify the name of the corpus (both its shorter name and its full descriptive name) and the location of the registry file (in this case: *In CQPweb's usual registry directory*), as highlighted in Figure 12.



Install a corpus pre-indexed in CWB

Click here to install a completely new corpus from files in the upload area.

Specify the corpus "name" (limited to lowercase/digits/underscore)
(will be used in the web address and as the CWB/MySQL short-handle)

Enter the full descriptive name of the corpus

Where is the registry file?
● In CQPweb's usual registry directory [?]
○ In the directory specified here:

Figure 12. *CQPweb Admin Control Panel (Install new corpus pre-indexed in CWB)*

[11] The alternative is *Click here to install a completely new corpus from files in the upload area*, but this was not applicable for the COVALT corpora.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Finally, we press *Install corpus with settings above*, located at the lower margin, as shown in Figure 13.

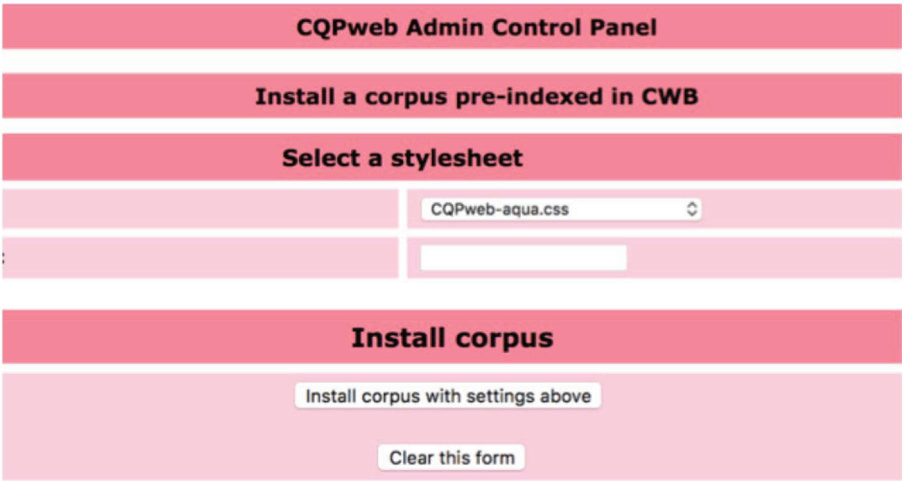

CQPweb Admin Control Panel
Install a corpus pre-indexed in CWB
Select a stylesheet
CQPweb-aqua.css
Install corpus
Install corpus with settings above
Clear this form

Figure 13. *CQPweb Admin Control Panel (Install corpus)*

Once the corpus has been installed in CQPweb, the next step is to generate the metadata (*Design and insert a text-metadata table for the corpus*). In the case of COVALT, the default option *Create minimalist metadata table* was chosen. This is the last step in the installation of monolingual corpora in CQPweb. For parallel corpora, there is an additional final step: *Manage parallel alignment*, located in the left margin.

Once all the steps detailed above have been completed, the subcorpora ANGESPOR and ANGESPTRAD will be available in CQPweb for online analysis.

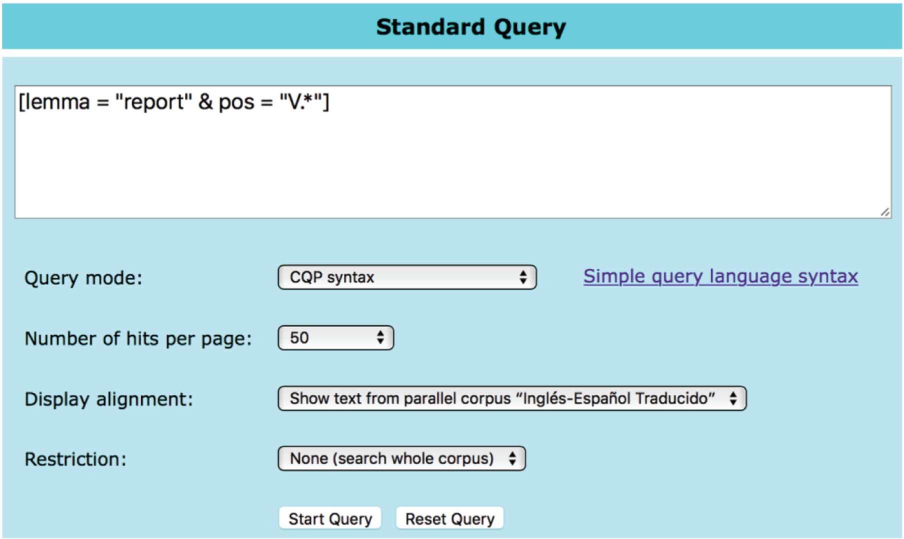
4. Corpus analysis

This section presents some example analyses of our parallel English – Spanish corpus, using the new version of CQPweb, IMS Open Corpus WorkBench's corpus analysis tool, which allows researchers to analyse corpora online.

In our example query, we search for the word *report* in the English source text corpus ANGESPOR. The options provided allow us to choose several query modes. One the one hand, to conduct a simple query (*report*). On the other hand, to use CQP syntax: a) [word = "report"], b) [lemma = "report"], c) [lemma = "report" &

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

pos = “NN”] or d) [lemma = “report” & pos = “V.*”].[12] The user can also determine the number of results per page (from 0 to 1000) and decide whether or not to display the aligned subcorpus on the results page.[13] Next an example query with the verb *report* by resorting to CQP syntax is illustrated (Figure 14).



[lemma = "report" & pos = "V.*"]

Query mode: CQP syntax

Number of hits per page: 50

Display alignment: Show text from parallel corpus “Inglés-Español Traducido”

Restriction: None (search whole corpus)

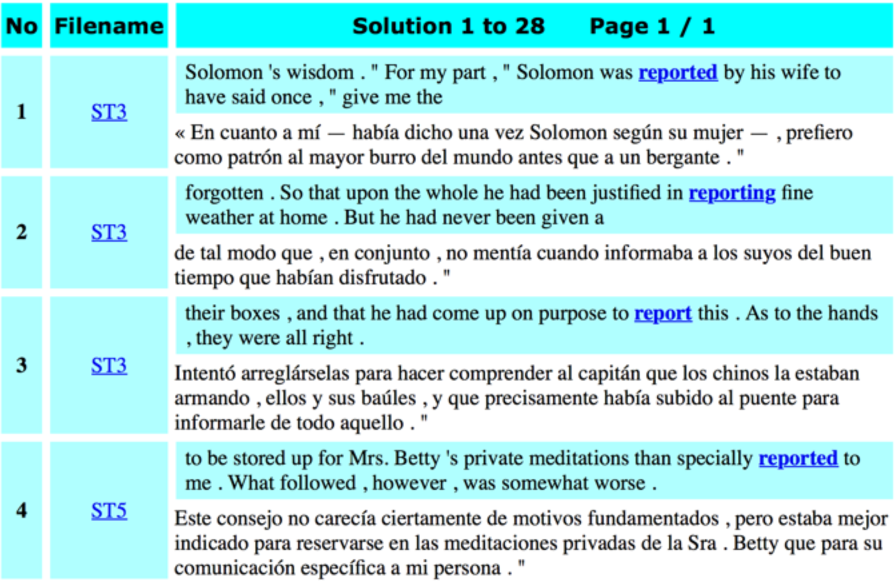
Figure 14. Query with CQP syntax

Figure 15 shows our results, with the corresponding aligned text fragments in Spanish. We also learn that there are 28 matches for the verb *report* in the corpus.

[12] In option *a* just the word *report* will be searched for; in option *b* all cases of the forms of the lemma *report* will be searched for; in option *c* all cases of the noun *report* will be searched for (“NN” is the tag for the grammatical category of noun in the English tagset used); in option *d* all cases of the verb *report* will be searched for (“V.*” is the tag for any kind of verb in the English tagset used).

[13] No restriction possibilities have currently been implemented.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



No	Filename	Solution 1 to 28 Page 1 / 1
1	ST3	Solomon 's wisdom . " For my part , " Solomon was reported by his wife to have said once , " give me the « En cuanto a mí — había dicho una vez Solomon según su mujer — , prefiero como patrón al mayor burro del mundo antes que a un bergante . "
2	ST3	forgotten . So that upon the whole he had been justified in reporting fine weather at home . But he had never been given a de tal modo que , en conjunto , no mentía cuando informaba a los suyos del buen tiempo que habían disfrutado . "
3	ST3	their boxes , and that he had come up on purpose to report this . As to the hands , they were all right . Intentó arreglárselas para hacer comprender al capitán que los chinos la estaban armando , ellos y sus baúles , y que precisamente había subido al puente para informarle de todo aquello . "
4	ST5	to be stored up for Mrs. Betty 's private meditations than specially reported to me . What followed , however , was somewhat worse . Este consejo no carecía ciertamente de motivos fundamentados , pero estaba mejor indicado para reservarse en las meditaciones privadas de la Sra . Betty que para su comunicación específica a mi persona . "

Figure 15. Query results (with aligned target text segments)

CQPweb also allows the researcher to perform specific actions without having to use CQP syntax. For instance, the user can save previous searches, sort the results or download them as a.txt document, or generate concordances, collocations, distribution tables and charts, frequency lists, etc. (cf. Hardie 2012). CQPweb includes a manual of *Simple query language syntax* with wildcards. Table 2 provides some examples.

Table 2. Simple query language syntax

Simple query language syntax	Results
s?ng	sing, sang, sung, song
*less	harmless, unless, helpless, etc.
black and white	all the concordances including this sequence
hono[u,]r	honor, honour

More complex searches, however, usually require CQP syntax. Table 3 presents some of the most important CQP search syntax options.[14]

[14] For a more detailed explanation, consult the CQP user's manual (Christ, Schulze, Hofmann, & König 1999) or the CQPWeb video tutorials <https://www.youtube.com/playlist?list=PL2XtJIhhrHNQgf4Dp6sckGZRU4NiUVw1e>.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 3. Main CQP syntax options

Action	CQP syntax
Search by grammatical category (e.g. proper noun)	`[pos = “NP”];`
Search by lemma	`[lemma = “be”];`
Count by word (applied to last search)	`count by word;`
Count by lemma (applied to last search)	`count by lemma;`
Search a sequence	

- lemma *go* + any word + preposition *to*
Example: […] Then she *went back to* the house […]
	`[lemma = “go”][]`
	`[word = “to”];`

- lemma *go* + two words + preposition *to* Example: […] and that is why I am *going to Oz to* ask him […]	`[lemma = “go”][] {2} [word = “to”];`

The Spanish section of the COVALT corpus has already been explored (by means of traditional CQP) in a series of studies, some of which have already been published. For instance, Oster & Molés-Cases (2016) have analysed the German>Spanish/Catalan subcorpora’s food-related culturemes, examining both eating and drinking as a source domain for metaphorical expressions and the expression of ways of eating and drinking. Molés-Cases (2016a) has also studied manner of motion events in a section of COVALT’s German>Spanish subcorpus (children’s and young adult literature).

5. Conclusion

This chapter has examined the process of compiling and indexing a parallel corpus on the CQPweb platform, using the COVALT English/French/German > Spanish parallel corpus as an example. Although the task of aligning, tagging and indexing a corpus to make it analysable online through CQPweb is a long-term project, the advantages the platform provides make the effort worthwhile. It offers an array of search options, a user-friendly interface and a means of formulating queries with the help of a simplified query language, instead of the more complex CQP syntax. Another crucial advantage of CQPweb is the key role of its online community. This includes, for instance, a highly participative mailing list (the CWB mailing list) run by the University of Bologna, which allows users to seek advice on problems with CWB and CQP and resolve difficulties speedily.

The Covalt research group aims to continue exploring and exploiting all four sections of COVALT for the analysis of translated Catalan and Spanish from a twofold perspective: through both the study of the parallel and the comparable subcorpora. This new web-based corpus analysis and management system will

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

provide the group with a more intuitive way of accessing and querying the corpora. Last but not least, CQPweb allows the Covalt group to share the resources it has developed with the wider research community.

Acknowledgements

We are grateful to Carme Colominas and her research group at the Universitat Pompeu Fabra who shared the Perl script *separab3.pl*, and to Daniel Renau for technical support on this project.

This study was supported by research projects FFI2015-68867-P, funded by the Spanish Ministry of the Economy and Competitiveness, and by P1-1B2013-44, funded by Universitat Jaume I.

References

Badia, Toni, Pujol, Manel, Tuells, Antoni, Vivaldi, Jorge, Yzaguirre, Lluís de & Cabré, Mª Teresa. 1998. IULA's LSP multilingual corpus: Compilation and processing. In *Proceedings of ELRA Conference*, 29–31 May 1998, Universidad de Granada.

Christ, Oliver, Schulze, Bruno, M., Hofmann, Anja & König, Esther. 1999. *The Open IMS Corpus Workbench. Corpus Query Processor. User's Manual*. Stuttgart: University of Stuttgart. <http://corpora.dslo.unibo.it/TCORIS/cqpman.pdf> (29 March 2017).

Doval, Irene, Fernández Lanza, Santiago, Jiménez Juliá, Tomás, Liste Lamas, Elsa & Lübke, Barbara. This volume. 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* [Studies in Corpus Linguistics 90], Irene Doval & M. Teresa Sánchez (eds). Amsterdam: John Benjamins.

Evert, Stefan & The CWB Development Team. 2016. *The IMS Open Corpus WorkBench (CWB). Corpus Encoding Tutorial*. <http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf> (20 October 2016).

Frankenberg-Garcia, Anna & Santos, Diana. 2003. Introducing Compara, the Portuguese English parallel corpus. In *Corpora in Translator Education*, Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds), 71–87. Manchester: St. Jerome.

Gómez Guinovart, Xavier & Sacau Fontela, Elena. 2004. Parallel corpora for the Galician Language: Building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 26–28 May 2004, Lisbon.

Guzman, Josep R. 2013. El corpus COVALT i l'eina d'alineament de frases Alfra-COVALT. In *El corpus COVALT: un observatori de fraseologia traduïda*, Llum Bracho Lapiedra (ed.), 49–60. Aachen: Shaker Verlag.

Guzman, Josep R. 2015a. Puntuació i traducció: *Verführung* i *Der Tangospieler*. *Quaderns – Revista de Traducció* 22: 217–232.

Guzman, Josep R. 2015b. Segmentation and regrouping of sentences. *Lenguaje y Textos* 42: 97–105.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Guzman, Josep R. 2016. La traducció de la modalitat deóntica i epistèmica del verb modal *sollen* en el corpus COVALT. *Zeitschrift für Katalanistik* 29: 135–165.

Hardie, Andrew. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409. <http://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf> (27 April 2017). https://doi.org/10.1075/ijcl.17.3.04har

Hardie, Andrew. 2014. *The IMS Open Corpus Workbench (CWB) CQPweb System Administrator's Manual*. <http://cwb.sourceforge.net/files/CQPwebAdminManual.pdf> (8 October 2016).

Hardie, Andrew. 2016. *The IMS Open Corpus Workbench (CWB). CQPweb System Administrator's Manual*. <http://cwb.sourceforge.net/> (20 November 2016).

Johansson, Stig. 2004. Multilingual corpora: models, methods, use. *TradTerm* 10: 59–82. https://doi.org/10.11606/issn.2317-9511.tradterm.2004.47044

Marco, Josep. 2013a. Tracing marked collocation in translated and non-translated literary language: A case study based on a parallel and comparable corpus. In *Tracks and Treks in Translation Studies* [Benjamins Translation Library 108], Catherine Way, Sonia Vandepitte, Reine Meylaerts & Magdalena Bartłomiejczyk (eds), 167–188. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.108.09mar

Marco, Josep. 2013b. La traducció de les unitats fraseològiques de base somàtica en el subcorpus angles català. In *El corpus COVALT: Un observatori de fraseologia traduïda*, Llum Bracho Lapiedra (ed.), 163–216. Aachen: Shaker Verlag.

Marco, Josep. 2018a The translation of food-related culture-specific items in the COVALT corpus: A study of techniques and factors. *Perspectives*. https://doi.org/10.1080/0907676X.2018.1449228.

Marco, Josep. 2018b Connectives as indicators of explicitation in literary translation: A study based on a comparable and parallel corpus. *Target* 30(1): 87–111. https://doi.org/10.1075/target.16042.mar

Martínez Vilinsky, Bárbara. 2016. La infrarrepresentación de elementos únicos en textos traducidos de ingles a español: perífrasis verbales, demostrativos y sufijos apreciativos en un corpus comparable y paralelo de novel policíaca. *PhD dissertation, Universitat Jaume I*, Castelló de la Plana, Spain.

Mikhailov, Mikhail & Cooper, Robert. 2016. *Corpus Linguistics for Translation and Contrastive Studies*. London: Routledge. https://doi.org/10.4324/9781315624570

Molés-Cases, Teresa. 2016a. *La traducción de los eventos de movimiento en un corpus paralelo alemán-español de literatura infantil y juvenil*. Frankfurt: Peter Lang. https://doi.org/10.3726/978-3-653-06745-3

Molés-Cases, Teresa. 2016b. Compilación y análisis de un corpus paralelo para la investigación en traducción. Proyecto con Déjà Vu, Treetagger e IMS Open Corpus Workbench. *RLA (Revista de Lingüística Teórica y Aplicada)* 54(1): 149–174. https://doi.org/10.4067/S0718-48832016000100008

Oksefjell, Signe. 1999. A description of the English – Norwegian parallel corpus: Compilation and further developments. *International Journal of Corpus Linguistics* 4(2): 197–219. https://doi.org/10.1075/ijcl.4.2.01oks

Oster, Ulrike & van Lawick, Heike. 2013. Anàlisi dels somatismes del subcorpus alemany-català. In *El corpus COVALT: Un observatori de fraseologia traduïda*, Llum Bracho Lapiedra (ed.), 267–294. Aachen: Shaker.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Oster, Ulrike & Molés-Cases, Teresa. 2016. Eating and drinking seen through translation: A study of food-related translation difficulties and techniques in a parallel corpus of literary texts. *Across Languages and Cultures* 17(1): 53–75. https://doi.org/10.1556/084.2016.17.1.3

Przepiórkowski, Adam, Górski, Rafał L., Łaziński, Marek & Pezik, Piotr. 2010. Recent developments in the National Corpus of Polish. In *Proceedings of the International Conference on Language Resources and Evaluation*, 17–23 May 2010, Valleta, Malta.

Sanjurjo-González, Hugo & Izquierdo, Marlen. This volume. P-ACTRES 2.0: A parallel corpus for cross-linguistic research. In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* [Studies in Corpus Linguistics 90], Irene Doval & M. Teresa Sánchez (eds). Amsterdam: John Benjamins.

Verdegal, Joan. 2013. Les unitats fraseològiqus somàtiques franceses i catalanes en COVALT: Localització, freqüència i anàlisi. In *El corpus COVALT: un observatori de fraseologia traduïda*, Llum Bracho Lapiedra (ed.), 217–266. Aachen: Shaker.

Verdegal, Joan. 2014. Traduir l'emoció: metodologia i resultats. In *Homenatge a Germà Colón. Labor omnia improbus vincit*, Rosa Agost & Lluís Gimeno (eds), 251–279. Castelló: Publicacions de la Universitat Jaume I.

Zubillaga, Naroa, Sanz, Zuriñe & Uribarri, Ibon. 2015. Building a trilingual parallel corpus to analyse literary translations from German into Basque. In *New Directions in Corpus-based Translation Studies*, Claudio Fantinuoli & Federico Zanettin (eds), 71–92. Berlin: Language Science Press.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

P-ACTRES 2.0

A parallel corpus for cross-linguistic research

Hugo Sanjurjo-González and Marlén Izquierdo
University of Huddersfield / University of the Basque Country (UPV/EHU)

This chapter describes an updated version of the ACTRES Parallel Corpus (P-ACTRES 2.0), an English-Spanish bidirectional corpus that contains over 4 million words. The composition of the corpus is recounted, regarding the number of words in each direction, and the types of texts included together with the linguistic variants that users will find in the corpus. Its composition is shaped by research purposes as well as availability issues. The computerization process is also explained, while commenting on the text processing, alignment and tagging. The chapter concludes with a brief demonstration of the usefulness and usability of P-ACTRES 2.0 in cross-linguistic research, be it contrastive linguistics or translation studies either independently or, most importantly, jointly.

Keywords: ACTRES, corpus analysis software, web interface, (parallel) corpus compilation

1. Introduction

The need for contrastive analyses that complement translation studies was proposed in the 1990s by scholars whose research was concerned with language applications that would improve cross-linguistic communication (Chesterman 1998; Rabadán 1991; Rabadán 2002b; Salkie 2002). Contrastive analysis and translation studies both originated within the branch of applied linguistics; yet, it took a while before both disciplines actually joined forces to undertake applications of language, in particular with regard to intercultural communication (Rabadán 1991). This was partly due to the fact that early cross-linguistic research lacked a functionalist perspective.

Prior to the development of descriptive translation studies (DTS), research focused mainly on the notion of (formal) equivalence understood as a sine qua non for translation. Confined to the realm of formal similarity, early contrastive

https://doi.org/10.1075/scl.90.13san
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

and translation research ignored units that, even though formally dissimilar, could actually communicate the same meanings. Possible differences in translation were not only overlooked, but also stigmatized (Venuti 2002). In early contrastive analysis (CA) studies pairs of languages were described parallel to but independently of one another without reference to cross-linguistic correspondences in discourse, even in the case of translations. Chesterman (1998, 2007) suggested that a way of overcoming these limitations would be to carry out contrastive functional analyses (CFA). Within this framework, the meaning we wish to translate, irrespective of the various forms it may be mapped upon, represents the tertium comparationis that makes contrasts possible. CFA, in turn, brought to light the one-to-many correspondences that may exist between two languages, which bears significant implications for translation. Furthermore, functionalism is sensitive to dissimilarities, which are highly relevant in cross-linguistic research since both sameness and difference constitute the (degree of) assumed similarity that is expected for any two languages to engage in an act of translation (Chesterman 2007).

In order to bridge the gap between the two, a functionalist approach to both contrastive linguistics and descriptive translation studies was called for. To this end, and drawing on functional grammar (Bondarko 1991), the ACTRES research team was founded in the second half of the 1990s, a time when new research tools were being developed (Rabadán 2002a). ACTRES stands for Análisis Contrastivo y TRaducción English–Spanish/ESpecializada . One of the benchmarks of functionalist linguistics in general and of ACTRES research in particular is the use of corpora (Johansson 1998). Early studies made use of existing reference corpora such as Bank of English by Cobuild and CREA (Corpus de Referencia del Español Actual). Comparability between these two corpora made it possible for ACTRES to carry out contrastive functional analyses using a lexicogrammar function such as quantification, noun characterization or modality as the object of study (Labrador de la Cruz 2000; Ramón García 2003; Rabadán 2006). However, when the starting point was a formal resource, such as the so-called gerund-participle (Huddleston & Pullum 2002) or -ing constructions, the custom-made comparable corpus based on Cobuild and CREA that had been used in these ACTRES studies turned out to be insufficient. This was so, given that many correspondents would remain unnoticed if English G-P were only compared to its assumed Spanish equivalent, namely the gerund (Izquierdo 2012) while other correspondences could be actually unveiled in authentic translations. It was because a truly joint contrastive translation study required a parallel corpus that the ACTRES team compiled P-ACTRES 1.0 (Izquierdo, Hofland & Reigem 2008).

This chapter describes P-ACTRES 2.0, an improved and extended version, which is by all accounts the first English–Spanish translation corpus ever built. P-ACTRES 2.0 is an open corpus in that new textual pairs of Spanish originals

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

and their English translations are being introduced as we speak. In this description of the corpus composition and compilation, the major differences between P-ACTRES 2.0 and the first, earlier version are highlighted. The final part of the description comprises a brief illustrative account of the kind of cross-linguistic research P-ACTRES 2.0 serves.

2. Description of the corpus

In this section, a thorough description of P-ACTRES 2.0 is provided. Given the fact that the corpus is an extension of P-ACTRES 1.0 (Izquierdo, Hofland & Reigem 2008), as each step of the corpus building is commented upon, a comparison of the two versions is made.

2.1 Composition

Composition represents a major distinguishing feature between the current and the earlier version. P-ACTRES 2.0 includes the P-ACTRES 1.0 composition and adds a new subcorpus composed of original Spanish texts and their translations into English. We will refer to the first group of texts as the English–Spanish (EN-ES) subcorpus or direction of analysis. The latter will be the Spanish–English (ES-EN) subcorpus, this being the major novelty in the composition of P-ACTRES 2.0.

With regard to the EN-ES subcorpus, original texts mainly represent British and American English, with a few samples of other varieties of English. The translations have been produced only in European Spanish. This decision was made so that the construction of P-ACTRES 1.0 assured the compatibility with previous ACTRES research based on comparable BoE-CREA. However, within the ES–EN direction we have included original texts written in both European and American Spanish. Regarding the target language varieties, both American and British English are found, which heightens the comparability between original English and translated English texts. The difference in the Spanish varieties per subcorpus is due to availability issues; on the one hand, Spanish translations of English texts outnumber English translations of Spanish texts, especially with regard to certain genres. Therefore, we needed to introduce original texts written in American Spanish to reach the corpus size we were aiming at. Also, related to availability, we wanted to benefit from the amount of ES–EN translations included in the UNESCO Index Translationum, a bibliographical database we used as a first step to locate potential pairs. In addition, another compilation criterion which places restrictions on the initial selection of texts was that publications should date back no earlier than the year 2000. Consequently, narrowing the target

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

variety to European Spanish only would have made the short-term compilation of P-ACTRES 2.0 virtually impossible.

In the first version, the corpus texts belong to five different genres, namely, fiction books, non-fiction books, newspaper editorials, magazine articles and miscellanea. With regard to the first two genres, the corpus contains excerpts rather than a whole book. The fact that not all the authors granted permission to include the whole book in the corpus – irrespective of the direction of analysis – forced us to use less than 10 per cent of every book so as to comply with copyright laws. In addition, the inclusion of excerpts rather than whole texts favoured the degree of language representativeness. In this sense, the regularity of patterns observed in the various studies carried out is most likely to represent the actual functioning of the languages rather than an idiosyncratic feature of a specific text or writer. Furthermore, having included different genres also accounts for representativeness keeping in mind that the corpus is compiled for research of non-specialized language. On the other hand, we would argue that the size of every genre mirrors how common its translation is, which also accounts for the representativeness and balance of P-ACTRES 1.0. In this regard, P-ACTRES 2.0 differs from the earlier version.

As hinted at above, availability has shaped the compilation of the ES-EN subcorpus so that no newspaper articles or miscellaneous texts have yet been compiled, as we have been unable to pinpoint genre-specific translations. Likewise, the sample of books-non-fiction is too small to preserve the balance of the ES–EN subcorpus.[1] Consequently, the compilation of P-ACTRES 2.0 remains ongoing to bridge these gaps and so, guarantee major representativeness and balance.

Table 1. Composition and size of P-ACTRES 2.0

English	Spanish	Total EN→ES	P-ACTRES 2,0	Total ES→EN	Spanish	English
1,276,791	1,357,296	2,634,087	Books-fiction 4,191,056	1,556,969	766,796	790,173
514,786	573,523	1,088,309	Books-nonfiction 1,116,582	28,273	15,068	13,205
118,665	129,810	248.475	Newspaper articles 248,475	–	–	–
114,634	122,024	236,658	Magazine articles 307,510	70,852	37,027	33,825
40,178	49,026	89,204	Miscellaneous 89,204	–	–	–
4,296,733 words			TOTAL 5,952,827 words	1,656,094 words		

[1] The sample size, as observed in Table 1, belongs to the material that is already browsable in the corpus. More text-pairs have already been aligned but the post-alignment computerization is still pending.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

By way of summary, Table 1 provides figures of the number of words per subcorpus (left and right columns) and of P-ACTRES 2.0 as a whole (central column). Note that earlier P-ACTRES 1.0 only covers the EN–ES translation direction.

2.2 Building of the corpus: Computational procedure

We will now explain the steps taken to computerize the corpus and make it easy to exploit technically.

2.2.1 *Technologies*

P-ACTRES 2.0 has been developed in a similar way as its predecessor P-ACTRES 1.0 (Izquierdo, Hofland & Reigem 2008). Several technologies as well as programming languages have been necessary in order to create an annotated bidirectional English–Spanish corpus, together with a user-friendly interface that enables effective corpus analysis. A brief description of the employed technologies is shown in Table 2.

Table 2. Employed technologies in the building of P-ACTRES 2.0

Type of technology	Use	Employed technology
General programming languages	Text processing and communication between user interface and data access layer.	Perl Bash
Web programming languages	Creation of user interface	HTML5 Javascript
External software	Text processing, aligning, corpus managing, searching, and linguistic tagging.	XML parsing programs such as: Xalan, Xerces or CleanXML TCA 2 (Translation Corpus Aligner, version 2) Treetagger The IMS Open Corpus Workbench (CWB)

The number of employed programming languages as well as external software serves as evidence of the complexity of the computational process carried out in order to build P-ACTRES 2.0. Several tasks had to be executed: text processing, aligning, corpus managing, corpus searching, linguistic tagging, communication between user interface and data access layer and user interface interaction among others. The most appropriate technology was used in order to complete the process adequately.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

2.2.2 *Workflow*

P-ACTRES 2.0 has been developed according to a very specific workflow divided into several stages, as shown in Figure 1.

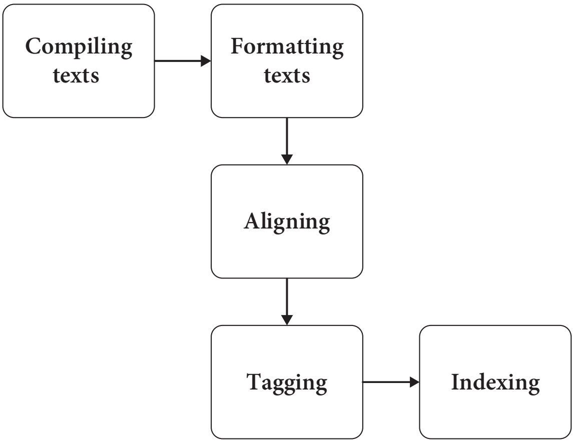


Figure 1. Workflow of the building of P-ACTRES 2.0

As mentioned above, P-ACTRES 2.0 is composed of two subcorpora that have not been compiled at the same time. Even though the workflow followed for the compilation of P-ACTRES 1.0 has been replicated in P-ACTRES 2.0, certain differences, in connection with compilation issues, are observed:

- Translations from Spanish texts into English are not as abundant as translations from English into Spanish.
- Increasingly, there is a major wealth of digitally processed data, so the scanning process has been required less and less in the compilation stage of P-ACTRES 2.0 (the ES–EN subcorpus).
- There are technical differences regarding indexing corpora. As the IMS Open Corpus Workbench (CWB) does not recognize comparable corpora, it is necessary to keep those subcorpora that are comparable through ad-hoc scripts. These differences also have an effect on querying the corpus (see Section 2.2.3).

i. *Compiling texts.* The first stage of the workflow is the compilation of the files that are part of the corpus. Original P-ACTRES 1.0 dates back to 2004; back at the time there were not as many resources on the Internet as we have at present. Moreover, the ACTRES team members were linguists whose computer literacy lacked programming skills. For this reason, most of the texts at that time were manually scanned, and thus, a proofreading process was necessary in order to clean noise made by the scanner (errors, misspellings, typos). The development of P-ACTRES 2.0 so far has involved the incorporation of 64 textual pairs, few of

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

which were already in digital format, so in these cases manual scanning and the ensuing proofreading process could be avoided.

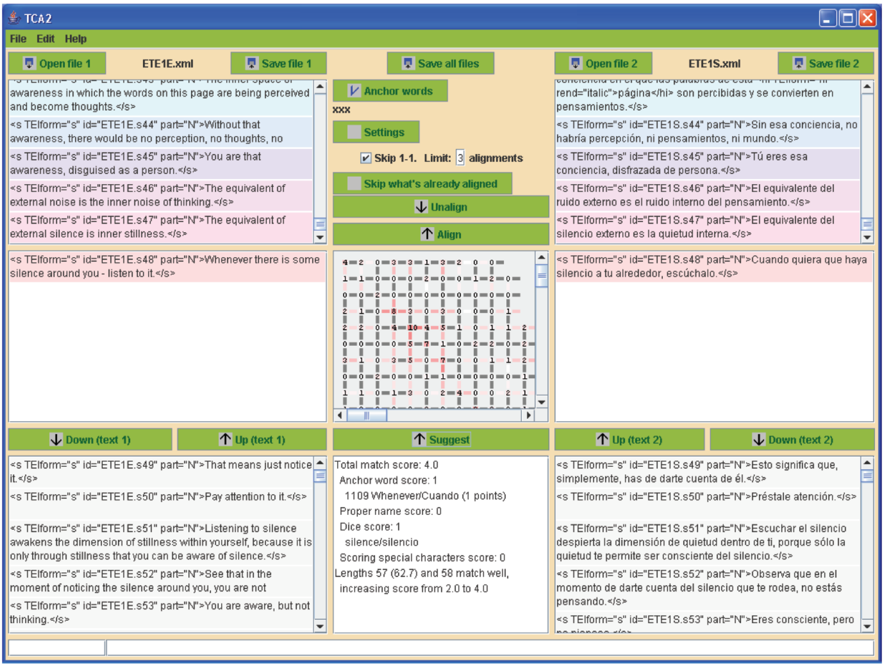
ii. *Formatting texts.* In this stage, each compiled document had to be formatted in order to be suitable for aligning and indexing processes. The formatting stage consists in:

- Cleaning texts: certain information had to be deleted – either multimodal or textual – in the texts so as to prevent noise in ensuing stages. This includes, for instance, footnotes, image references or captions, tables, page numbers and non-valid characters, among others.
- Transforming texts into XML format: all the texts that composed the corpus have been converted into XML data format. This data format allows us to handle the documents in an efficient way thanks to its standarized structure.
- Validating XML files: this process ensures that features of the texts are equally structured. For instance, titles, sections or metadata.
- Carrying out a sentence division: each sentence of the text is identified and assigned a unique ID that recognises the sentence as belonging to a specific file. This is essential for the alignment stage.

iii. *Aligning the textual pairs.* The alignment of the corpus represented a highly time-consuming stage. It was carried out using a second version of the Translation Corpus Aligner program or TCA (Hofland & Johansson 1994), originally developed by Knut Hofland for the English–Norwegian Parallel Corpus (ENPC), one of the first parallel, aligned corpus of all time (ibid). TCA2 had previously been improved by Knut Hofland and Øystein Reigem for P-ACTRES 1.0; it is an unsupervized, language-independent alignment program at the sentence level which allows manual correction.

TCA2 has a high accuracy rate of 95–97% over 1 to 1 alignments, that is, one sentence of original text is aligned with one sentence of translated text. However, manual correction is needed to get a perfect alignment in n-1 or 1-n alignments. This is the reason why a semi-automatic alignment procedure has been adopted in both versions, that is, P-ACTRES 1.0 and P-ACTRES 2.0, instead of running the program automatically on all the files. Figure 2 shows the alignment of a textual pair.

By default, the program favours 1-1 alignments, which is the reason why the user has the possibility of selecting the "skip 1-1" function inbuilt in the TCA2. Consequently, if the program suggests a correspondence other than 1-1, that is 1-n, n-1 or even a null correspondence, it will stop for the user to check whether that suggestion is correct or not. In the event of an incorrect match, the user has the possibility of clicking the "unalign" bottom. Whereas, if there actually is a



```
awareness in which the words on this page are being perceived
and become thoughts.</s>
<s TEIform="s" id="ETE1E.s44" part="N">Without that
awareness, there would be no perception, no thoughts, no
<s TEIform="s" id="ETE1E.s45" part="N">You are that
awareness, disguised as a person.</s>
<s TEIform="s" id="ETE1E.s46" part="N">The equivalent of
external noise is the inner noise of thinking.</s>
<s TEIform="s" id="ETE1E.s47" part="N">The equivalent of
external silence is inner stillness.</s>

<s TEIform="s" id="ETE1E.s48" part="N">Whenever there is some
silence around you - listen to it.</s>

<s TEIform="s" id="ETE1E.s49" part="N">That means just notice
it.</s>
<s TEIform="s" id="ETE1E.s50" part="N">Pay attention to it.</s>
<s TEIform="s" id="ETE1E.s51" part="N">Listening to silence
awakens the dimension of stillness within yourself, because it is
only through stillness that you can be aware of silence.</s>
<s TEIform="s" id="ETE1E.s52" part="N">See that in the
moment of noticing the silence around you, you are not
<s TEIform="s" id="ETE1E.s53" part="N">You are aware, but not
thinking.</s>
```

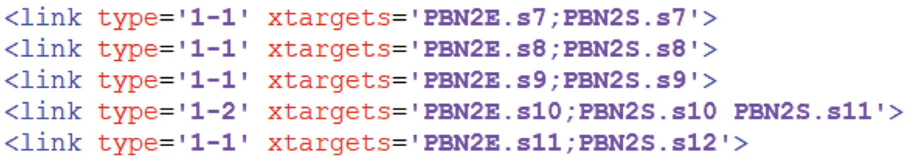
```
Total match score: 4.0
 Anchor word score: 1
  1109 Whenever/Cuando (1 points)
 Proper name score: 0
 Dice score: 1
  silence/silencio
 Scoring special characters score: 0
Lengths 57 (62.7) and 58 match well,
 increasing score from 2.0 to 4.0
```

```
rend="italic">página</hi> son percibidas y se convierten en
pensamientos.</s>
<s TEIform="s" id="ETE1S.s44" part="N">Sin esa conciencia, no
habría percepción, ni pensamientos, ni mundo.</s>
<s TEIform="s" id="ETE1S.s45" part="N">Tú eres esa
conciencia, disfrazada de persona.</s>
<s TEIform="s" id="ETE1S.s46" part="N">El equivalente del
ruido externo es el ruido interno del pensamiento.</s>
<s TEIform="s" id="ETE1S.s47" part="N">El equivalente del
silencio externo es la quietud interna.</s>

<s TEIform="s" id="ETE1S.s48" part="N">Cuando quiera que haya
silencio a tu alrededor, escúchalo.</s>

<s TEIform="s" id="ETE1S.s49" part="N">Esto significa que,
simplemente, has de darte cuenta de él.</s>
<s TEIform="s" id="ETE1S.s50" part="N">Préstale atención.</s>
<s TEIform="s" id="ETE1S.s51" part="N">Escuchar el silencio
despierta la dimensión de quietud dentro de ti, porque sólo la
quietud te permite ser consciente del silencio.</s>
<s TEIform="s" id="ETE1S.s52" part="N">Observa que en el
momento de darte cuenta del silencio que te rodea, no estás
pensando.</s>
<s TEIform="s" id="ETE1S.s53" part="N">Eres consciente, pero
```

Figure 2. Image of TCA2 interface

multiple sentence match, the user will click on align for approval. A colour-code in the upper windows helps to identify the aligned matches or correspondences.

The output of the program is an alignment file (Figure 3) and text files. The former is composed of several lines that are made up of XML elements. These elements contain information about the type of correspondence (type) and the involved sentences through their identification (xtargets).



```
<link type='1-1' xtargets='PBN2E.s7;PBN2S.s7'>
<link type='1-1' xtargets='PBN2E.s8;PBN2S.s8'>
<link type='1-1' xtargets='PBN2E.s9;PBN2S.s9'>
<link type='1-2' xtargets='PBN2E.s10;PBN2S.s10 PBN2S.s11'>
<link type='1-1' xtargets='PBN2E.s11;PBN2S.s12'>
```

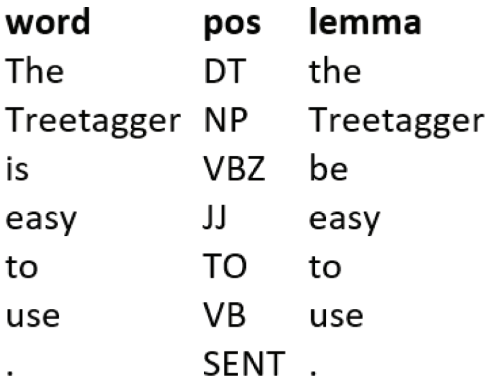
Figure 3. Image of an alignment file

iv. *Tagging.* The next stage consists in adding linguistic annotation to the corpus. P-ACTRES 2.0 has been tagged with grammatical and lemma information using Treetagger (Schmid 1995)[2]; This software is freely accessible online and has

[2] For more information about the employed tagsets, see: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf> for English and <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Spanish–tagset.txt> for Spanish.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

been used in a great variety of corpora of different languages. Figure 4 shows an example of the output of Treetagger.



```
word        pos   lemma
The         DT    the
Treetagger  NP    Treetagger
is          VBZ   be
easy        JJ    easy
to          TO    to
use         VB    use
.           SENT  .
```

Figure 4. Image of a POS tagged sentence

v. *Indexing.* The last building step consists in executing several programming scripts in order to format the corpus files according to the requirements of the IMS Corpus Workbench (from now CWB) (Evert & Hardie 2011), which has been the corpus management software chosen. The aforementioned requirements of CWB are the following:

- One file per corpus, which means that all the text files had to be put together in one file.
- The corpus file had to be in verticalized format, which is one word per line, and linguistic annotations separated by tab values. Sentences and features of the texts have to be handled according to XML syntax.

CWB allows for querying the corpus and works as a data access layer of P-ACTRES 2.0. For more info about corpus indexing and querying with CWB, see Evert (2016) and Evert (2009).

CWB manages P-ACTRES 2.0 as if it were composed of four different subcorpora or text groups:

- Original English texts
- Spanish translations
- Original Spanish texts
- English translations.

While CWB recognizes the parallel concordances from original English into Spanish as well as original Spanish into English, it was not originally designed for comparable searches. As a consequence, for the comparable exploitation of the corpus, that is, keeping those subcorpora that are comparable for doing searches on original and translated texts per language,[3] ad-hoc scripts had to be written. This step is further explained in Section 2.2.3, where the browser or user interface is described.

2.2.3 *User interface*

The combination of two directions of analysis extends the functionality of P-ACTRES 2.0 compared with P-ACTRES 1.0, which only allowed us to query the English – Spanish parallel subcorpus. P-ACTRES 2.0 provides us with the possibility of querying over the following subcorpora:

- English–Spanish Parallel corpus, former P-ACTRES 1.0;
- Spanish–English Parallel corpus;
- English–Spanish Comparable corpus (both original);
- Translation corpus of original and translated English;
- Translation corpus of original and translated Spanish;
- Even the study of the so-called "third code", that is, only translated language (English and Spanish translated texts).

The user interface of P-ACTRES 2.0 has been completely renewed, providing information in a colourful and visually appealing format. Figure 5 shows a comparison between both interfaces; element (i) shows the previous P-ACTRES 1.0 interface, whereas element (ii) shows the new interface for parallel queries, and element (iii) for comparable queries.

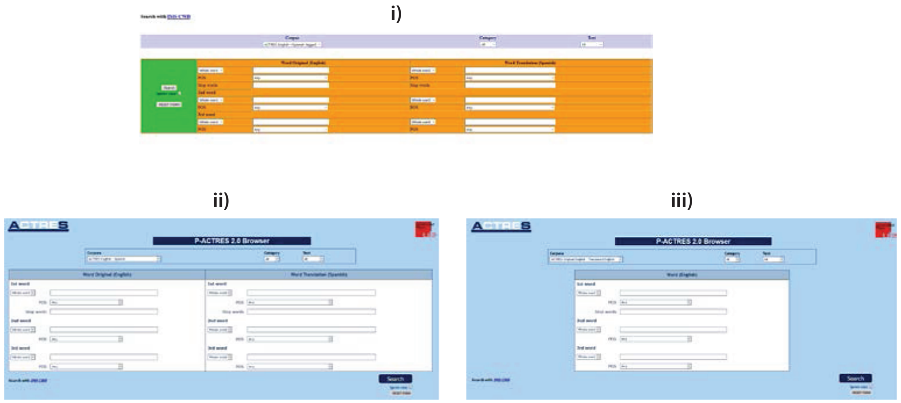
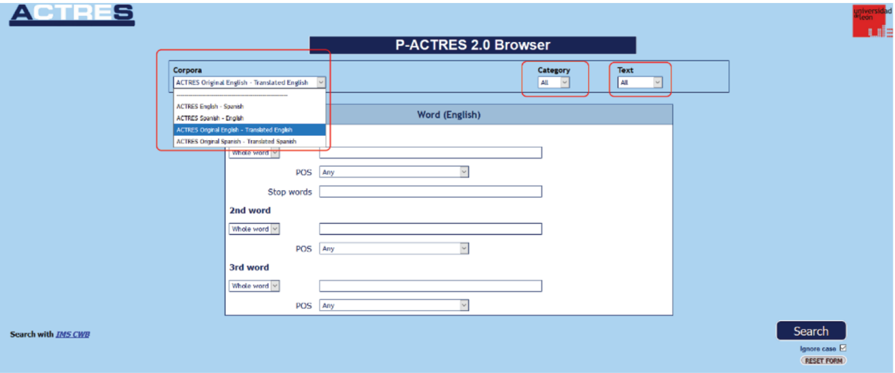


Figure 5. User interfaces comparison

[3] Comparable queries that involve only the original texts per language are fully supported by CWB.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Figure 6 shows an image of the current browser interface and how to select the desired subcorpus, the genre of the texts or even a determined text.



Corpora: ACTRES Original English - Translated English
Category: All
Text: All
Whole word
POS: Any
Stop words
Whole word
POS: Any
Whole word
POS: Any
Ignore case ☑

Figure 6. Corpus selection in P-ACTRES 2.0 browser

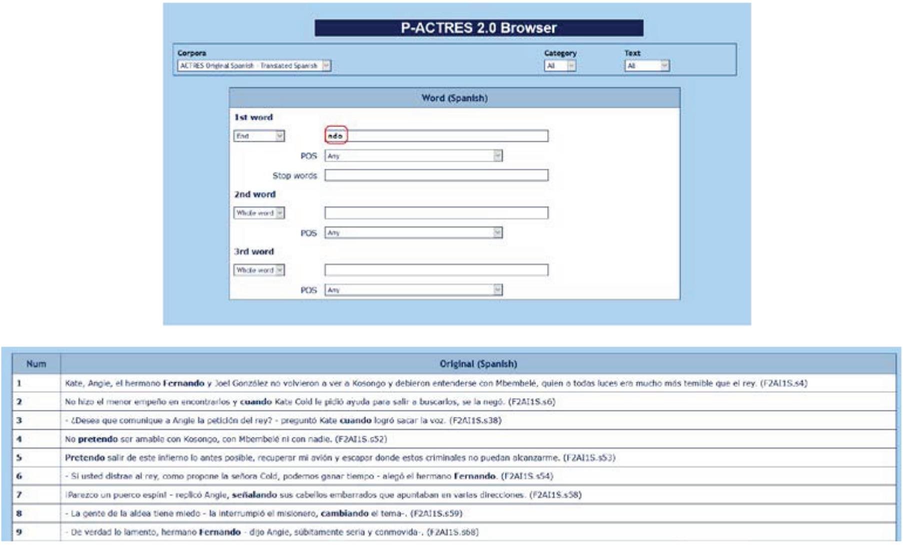
P-ACTRES 2.0 uses CWB to take advantage of CQP Query Language or CQP-syntax (Evert 2009) in order to execute queries over large annotated corpora, whether they are monolingual or parallel. CWB, however, does not support comparable queries natively, which is a new functionality of P-ACTRES 2.0, so several programming scripts have had to be executed in the background to make comparable corpora a reality. This consists basically in the simultaneous execution of the same query in two subcorpora that are considered comparable.

In order to simplify querying process, P-ACTRES 2.0 has been equipped with the same search system as P-ACTRES 1.0, but providing new interfaces for comparable queries. All the interfaces are similar and consist of several input fields that let the user define the type of sequence to be searched, as shown above. In this way, search patterns of CQP Query Language are replaced by simple instructions, as explained in (1):

(1) If you want to search all the words that start with a specific pattern, which in CQP Query Language is, for instance, [word="beaut.*"]), you only have to select the option "Start".

Depending on the type of search, the browser provides us with several options (see Figure 7): a whole word, a set POS tag, the beginning of a word or its ending. The user could even specify stop words, that is, unwanted words the corpus is asked to skip and not to retrieve. Finally, if the query were a multi-word unit or string of words, the beginning of the sequence or POS constituents may be specified, among other options (see Rabadán in this volume: Section 2, Figure 4 for POS options; Section 5, Figure 5 for aligned results).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use


P-ACTRES 2.0 Browser

Corpora: ACTRES Original Spanish - Translated Spanish
Category: All
Text: All

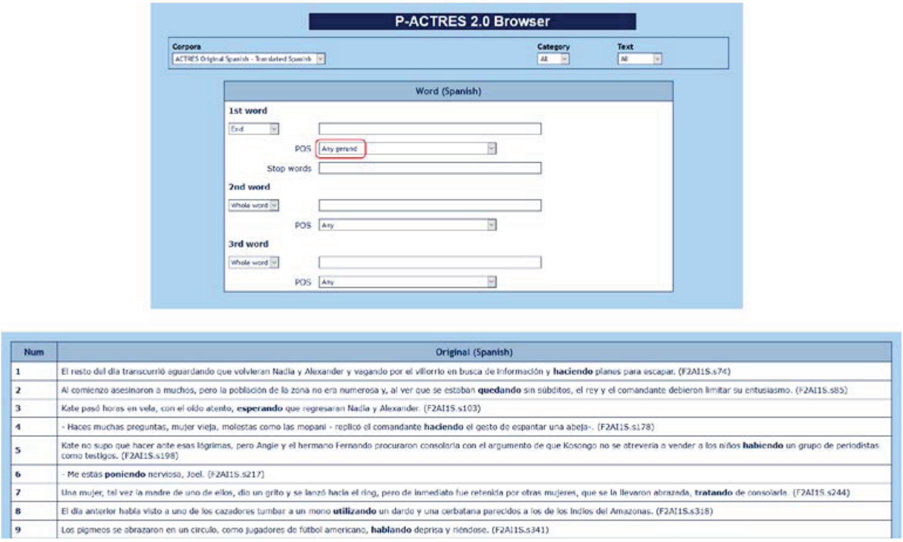
Word (Spanish)

1st word: End — ndo
POS: Any
Stop words:
2nd word: Whole word
POS: Any
3rd word: Whole word
POS: Any

Num	Original (Spanish)
1	Kate, Angie, el hermano **Fernando** y Joel González no volvieron a ver a Kosongo y debieron entenderse con Mbembelé, quien a todas luces era mucho más temible que el rey. (F2AI1S.s4)
2	No hizo el menor empeño en encontrarlos y **cuando** Kate Cold le pidió ayuda para salir a buscarlos, se la negó. (F2AI1S.s6)
3	- ¿Desea que comunique a Angie la petición del rey? - preguntó Kate **cuando** logró sacar la voz. (F2AI1S.s38)
4	No **pretendo** ser amable con Kosongo, con Mbembelé ni con nadie. (F2AI1S.s52)
5	**Pretendo** salir de este infierno lo antes posible, recuperar mi avión y escapar donde estos criminales no puedan alcanzarme. (F2AI1S.s53)
6	- Si usted distrae al rey, como propone la señora Cold, podemos ganar tiempo - alegó el hermano **Fernando**. (F2AI1S.s54)
7	¡Parezco un puerco espín! - replicó Angie, **señalando** sus cabellos embarrados que apuntaban en varias direcciones. (F2AI1S.s58)
8	- La gente de la aldea tiene miedo - la interrumpió el misionero, **cambiando** el tema-. (F2AI1S.s59)
9	- De verdad lo lamento, hermano **Fernando** - dijo Angie, súbitamente seria y conmovida-. (F2AI1S.s68)

Figure 7. Search with query "any word ending in -ndo" and results

Imagine the user wanted to identify gerund constructions in Spanish. One option would be to ask the corpus to search for words ending in -ando. However, this pattern will automatically skip gerund forms in -endo (*leyendo, diciendo*, etc.). We could, therefore, specify the word ending as -ndo. Nonetheless, with such an option we run the risk of getting noise: proper names such as "Fernando", non-gerund verbs like *pretendo* or particles such as *cuando*, among others (Figure 7).


P-ACTRES 2.0 Browser

Corpora: ACTRES Original Spanish - Translated Spanish
Category: All
Text: All

Word (Spanish)

1st word: End
POS: Any gerund
Stop words:
2nd word: Whole word
POS: Any
3rd word: Whole word
POS: Any

Num	Original (Spanish)
1	El resto del día transcurrió aguardando que volvieran Nadia y Alexander y vagando por el villorrio en busca de información y **haciendo** planes para escapar. (F2AI1S.s74)
2	Al comienzo asesinaron a muchos, pero la población de la zona no era numerosa y, al ver que se estaban **quedando** sin súbditos, el rey y el comandante debieron limitar su entusiasmo. (F2AI1S.s85)
3	Kate pasó horas en vela, con el oído atento, **esperando** que regresaran Nadia y Alexander. (F2AI1S.s103)
4	- Haces muchas preguntas, mujer vieja, molestas como las mopani - replicó el comandante **haciendo** el gesto de espantar una abeja-. (F2AI1S.s178)
5	Kate no supo qué hacer ante esas lágrimas, pero Angie y el hermano Fernando procuraron consolarla con el argumento de que Kosongo no se atrevería a vender a los niños **habiendo** un grupo de periodistas como testigos. (F2AI1S.s198)
6	- Me estás **poniendo** nerviosa, Joel. (F2AI1S.s217)
7	Una mujer, tal vez la madre de uno de ellos, dio un grito y se lanzó hacia el ring, pero de inmediato fue retenida por otras mujeres, que se la llevaron abrazada, **tratando** de consolarla. (F2AI1S.s244)
8	El día anterior había visto a uno de los cazadores tumbar a un mono **utilizando** un dardo y una cerbatana parecidos a los de los indios del Amazonas. (F2AI1S.s318)
9	Los pigmeos se abrazaron en un círculo, como jugadores de fútbol americano, **hablando** deprisa y riéndose. (F2AI1S.s341)

Figure 8. Search with query "any G" and results

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

An alternative search would be to ask the corpus to retrieve any word POS tagged as G, instead (Figure 8).

2.3 P-ACTRES 1.0 VS P-ACTRES 2.0

As the extension of P-ACTRES 1.0, the building of P-ACTRES 2.0 has been explained in the sections above, we have referred to the major differences between the two versions. Table 3 provides the reader with a straightforward summary of these major differences.

Table 3. Differences between P-ACTRES 1.0 and P-ACTRES 2.0

P-ACTRES 1.0	**P-ACTRES 2.0**	**NOVELTIES in 2.0**
2,523,458 words	5,952,827 words	Makes corpus larger
Unidirectional: English -> Spanish	Bidirectional English <-> Spanish	New direction of analysis (ES–EN)
5 subcorpora: Fiction Scientific/Academic writing Newspaper articles Magazine articles Miscellaneous	5 subcorpora (En->Sp): Fiction Scientific/Academic writing Newspaper articles Magazine articles Miscellaneous 3 subcorpora (Sp->En): Fiction Scientific/Academic writing Magazine articles	New textual pairs
Queries over parallel corpora	Queries over parallel corpora Queries over comparable corpora	New browsing possibilities More textual combinations for research
One simple search interface	Two user-friendly search interfaces	New browser interface

3. Usefulness and usability in cross-linguistic research

In this section, we will show how P-ACTRES 2.0 may be exploited in cross-linguistic research, especially in a descriptive stage. The research possibilities in terms of language direction and type of language use (authentic and/or translated) resemble those of the English–Norwegian Parallel Corpus (ENPC) (Johansson & Hofland 1994). As it was the first parallel corpus of sentence-aligned translations, the ENPC has inspired many other corpora (ESPC, OMS, to name but a few).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Figure 9 below shows the various ways in which the corpus material may be browsed:

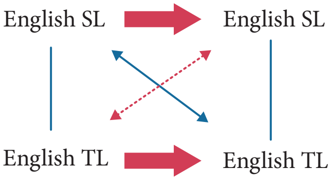


Figure 9. Browsing options

Option 1: parallel-corpus-based translation studies
(a) English OT > Spanish TT
or
(b) Spanish OT > English TT

Option 2: comparable-corpus-based contrastive studies
(c) English OT + Spanish OT

Option 3: corpus-based joint contrastive and translation studies
(a) + (b) and (a) + (b) + (c) above

Option 4: translation studies
(d) Original English + Translated English
(e) Original Spanish + Translated Spanish

Option 5: translation-corpus-based studies of the third code
(f) Translated English + Translated Spanish

While any of the options above are useful for cross-linguistic research, it is option 3 that truly reflects the ACTRES research endeavour. As will be shown, the combination of parallel data with comparable data will reveal functional equivalents that would be either impossible to track down or would be mismatched on the basis of formal similarity alone. By way of example we take as our object of study a specific lexicogrammatical construction of English, namely, "with + NP + -ing" (as, for example, "…, **with ministries being turned** over nearly every day now as …"); An analysis of this sequence in context reveals its adverbial function, that is, elaborating on something that has been said previously. From a functional point of view and taking the –ing form or gerund-participle (G-P) as its main constituent, this construction belongs in so-called G-P Adjuncts (Izquierdo 2012), a non-finite subordinate type clause.

Two questions may trigger our cross-linguistic research: first, from a contrastive analysis point of view, if the Spanish Gerund (G) is the assumed equivalent of the English G-P, does the Spanish resource take part in a similar construction?

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

In other words, does Spanish feature constructions of the kind "….con los últimos siendo los más frecuentes…"? Second, from the point of view of translation studies, if a formally similar construction exists in Spanish, is this the most recurrently used equivalent? But what happens if there is not a formally similar lexicogrammatical resource in Spanish? What other translational options should be used instead? (see Table 4).

Table 4. Spanish translational options of English non-finite construction "with + NP + G-P"

English ST	**Spanish TT**	**N**	**%**
…(,) with N -ing	Prepositional Phrase (con)	11	32.3
	Subordinate –finite- clause	10	29.4
	Adverbial (7)Nominal (2) Relative (1)		
	Coordinate clause	9	26.5
	Others	3	8.8
	No translation	1	3
	TOTAL	34	100

Drawing from P-ACTRES comparable data, that is, original language use, we observe that the Spanish G also takes on an adverbial function at the sentence level. We refer to this non-finite clause as G Adjunct. However, whereas the formal realization of the English G-P Adjunct allows for introductory particles and its own subject, this is not the case in Spanish (Izquierdo 2012). Furthermore, a construction of the kind "con + NP + G" would be considered unidiomatic in Spanish given that, according to comparable data, the G does not display adjectival or predicative functions after NPs.[4] So, how should we translate these constructions into Spanish? According to Malmkjaer (2005: 15), "[equivalence is] the relationship which actually obtains between the translation and the source text: an empirical, rather than an ideal phenomenon, open to description." Therefore, let us examine actual translations to identify functional correspondences. In other words, let us resort to a parallel corpus so as to bridge the gap identified in the previous contrastive stage.

Introducing the query string "with + any N + ending -ing", the corpus retrieves 62 hits, 34 of which are valid. The remaining constructions are considered noise because the G-P is actually either a fossilized noun or an adjective, as in "…four of us with part-time cleaning women from Reydon …".

[4] The exeption being NPs in verbal complements required by certain types of verbs like "ver + noun + G" (see + N + -ing).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Therefore, we examine the translation of the 34 valid parallel concordances to find out the translational options (see Table 4):

The most frequently used translational option is a prepositional phrase (PP) introduced by “con” (with). This equivalent conveys a downrank shift whereby the TT occurs below the sentence level, as unlike in the ST, it is not a clause (Example 1).

(1) Ex 1. PP
TO: Data emerging from Continental Europe was mixed, ***with figures confirming*** a sharp rise in confidence and output expectations and leading indicators hinting at a peaking growth cycle. (M23E.s426)
TM: Los datos procedentes de la Europa continental eran heterogéneos, ***con cifras que confirmaban*** un veloz aumento de la confianza y de las expectativas de resultados y los principales indicadores mostraban que el ciclo de crecimientos estaba en su punto álgido. (M23S.s421)

This similarity is maintained, however, with the second equivalent in frequency of use, namely a finite subordinate clause. As expected, the majority of this type of clause carries adverbial meanings, as observed in Example 2.

(2) Ex 2. Finite, subordinate clause
TO: But it seems to be a sprint to the finish, ***with ministries being turned*** over nearly every day now as American officials take an increasingly low profile. (PGEJ1E.s57)
TM: Casi a diario se entregan ministerios a los iraquíes, ***mientras que los estadounidenses asumen*** un papel cada vez más secundario. (PGEJ1S.s46)

The other translational options do not keep the adverbial meaning of the original clause. Consequently, even when they are correct in Spanish, their degree of equivalence is (arguably) really low.

4. Conclusions

The building of the ACTRES Parallel Corpus (both versions) has turned out to be a very complex and laborious process that involved several technologies. It was due to the limited resources in both bilingual parallel corpora and comparable corpora featuring the Spanish language that the creation of P-ACTRES 2.0 became a priority for the ACTRES research group in their endeavours to carry out effective corpus-based contrastive studies and translation studies whether independently or jointly. The steps described are part of a habitual or mainstream procedure to build most corpora.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Given the scope of this chapter, we have merely outlined the fruitful alliance such parallel corpora offer contrastive linguists and translation scholars. While contrastivists can benefit from translation studies that unveil correspondents that would a priori be impossible to conceive, researchers in translation studies can capitalize on contrastive work using functional correspondents that may work as translation equivalents. In other words, while examining translations can reveal objects of contrastive study, descriptive work within contrastive linguistics will provide guidelines or tools for use in applications of language such as translation.

Further work on the P-ACTRES 2.0 features is needed; for example, new annotation layers, such as all-words tagging semantic annotation or composition/ rhetorical annotation, will prove useful. Also included in plans for the near future are the implementation of more sophisticated statistics such as graphic charts or statistics based on significance tests, based on the flexibility and power of the statistics-oriented programming language R.

Acknowledgements

Research for this chapter, undertaken as part of the ACTRES program, has been partly funded by the Spanish Ministry of Science and Innovation and ERDF [FFI2016-75672-R]. The acronym stands for Análisis contrastivo y traducción English–Spanish "Contrastive analysis and translation English–Spanish".

References

Chesterman, Andrew. 1998. *Contrastive Functional Analysis* [Pragmatics & Beyond New Series 47]. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.47

Chesterman, Andrew. 2007. Similarity analysis and the translation profile. In *The Study of Language and Translation* [Belgian Journal of Linguistics 21], Willy Vandeweghe, Sonia Vandepitte & Marc Van de Velde (eds), 53-66. Amsterdam: John Benjamins. https://doi.org/10.1075/bjl.21.05che

Evert, Stefan. 2009. *The CQP Query Language Tutorial*. <http://cwb.sourceforge.net/files/CQP_Tutorial.pdf> (10 May 2017).

Evert, Stefan. 2016. *The IMS Open Corpus Workbench (CWB) – Corpus Encoding Tutorial*. <http://cwb.sourceforge.net/files/CQP_Tutorial.pdf> (10 May 2017).

Evert, Stefan & Hardie, Andrew. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics Conference 2011*. Birmingham: University of Birmingham.

Hofland, Knut & Johansson, Stig. 1998. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. *Language and Computers* 24: 87–100.

Huddleston, Rodney & Pullum, Geoff. 2002. *The Cambridge Grammar of the English Language*. Cambridge: CUP.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Izquierdo, Marlén. 2012. *Estudio contrastivo y de traducción inglés-español*. Saarbrücken: Editorial Académica Española.

Izquierdo, Marlén, Hofland, Knut & Reigem, Øystein. 2008. The ACTRES parallel corpus: An English–Spanish translation corpus. *Corpora* 3(1): 31–41. https://doi.org/10.3366/E1749503208000051

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

Johansson, Stig. & Hofland, Knut. 1994. Towards an English–Norwegian parallel corpus. In *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zurich 1993*, Udo Fries, Gunnel Tottie & Peter Schneider (eds), 25–37. Amsterdam: Rodopi.

Labrador de la Cruz, Belén. [2000]2005. *Estudio contrastivo de la cuantificación inglés-español*. León: University of León.

Malmkjaer, Kirsten. 2005. *Linguistics and the Language of Translation*. Edinburgh: EUP.

Rabadán, Rosa. 2006. Modality and modal verbs in contrast: Mapping out a translation(ally) relevant approach English–Spanish. *Languages in Contrast* 6(2): 261–306. https://doi.org/10.1075/lic.6.2.04rab

Rabadán, Rosa. 2002a. Análisis contrastivo y traducción inglés-español: El programa ACTRES. In *Nuevas perspectivas de los estudios de traducción*, Jose Maria Bravo & Purificación Fernández (eds), 35–55. Valladolid: University of Valladolid.

Rabadán, Rosa. 2002b. Normativity and functionality in translation English–Spanish: Theory and contrast. In *Actas del XXVI Congreso de AEDEAN*, 725–733. Granada: University of Granada.

Rabadán, Rosa. 1991. *Equivalencia y traducción. Problemática de la equivalencia translémica inglés-español*. León: University of León.

Ramón, Noelia. 2003. *Estudio contrastivo inglés-español de la caracterización de sustantivos*. León: University of León.

Salkie, Raphael. 2002. How can linguists profit from parallel corpora? In *Parallel Corpora, Parallel Worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April 1999*, Lars Borin (ed.), 93–109. Amsterdam: Rodopi.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In *Natural Language Processing Using Very Large Corpora*, Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann & David Yarowsky (eds), 13–25. Dordrecht: Springer.

Venuti, Lawrence. 2002. The difference than translation makes: the translator's unconscious. In *Translation Studies: Perspectives on an Emerging Discipline*, Alessandra Riccardi (ed.), 214–241. Cambridge: CUP.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

An overview of Basque corpora and the extraction of certain multi-word expressions from a translational corpus

Zuriñe Sanz-Villar
University of the Basque Country (UPV/EHU)

Since the 1980s, considerable efforts have been made to create different types of Basque corpora. However, to systematically analyse the Basque translations of German literary texts, it was necessary to create a corpus from the ground up. Intermediary versions were included in this corpus whenever the Basque target text was not a translation from the German original but came instead from a translation into another language (Spanish in most cases). A tool called TAligner was used to align the bitexts and the tritexts. The aim of this chapter is, firstly, to provide the reader with an overview of the main Basque corpora. Secondly, I will describe the design and compilation process of a parallel and multilingual corpus using TAligner 3.0. Thirdly, I will present how the corpus has been lemmatized and annotated at the level of part-of-speech. Finally, the process of extracting potential Basque multi-word expressions will be shown.

Keywords: Basque corpora, Aleuska corpus, TAligner, Basque MWEs

1. Introduction

According to Sinclair (2005), a corpus is "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language variety as a source of data for linguistic research". The first electronic corpus, the *Brown Corpus*, was compiled in 1967 at Brown University by Kučera and Francis. Although a number of other attempts were made in the following years, few corpora were successfully compiled before the 1980s. Zanettin (2012: 7) mentions two factors that have contributed to this deficiency: technological restrictions, and the Chomskyan or mentalist approach, which was in vogue in the United States during the second half of the twentieth century.

https://doi.org/10.1075/scl.90.14san

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

© 2019 John Benjamins Publishing Company

The prevalence of the Chomskyan paradigm, where the focus lay on how languages are structured in the human brain (Kenny 2001: 3) rather than on analysing the actual use of languages, hindered scholars wishing to work with corpora. Later, corpus linguistics provided answers to some of the questions that the Chomskyan approach could not. This led to the development of corpus linguistics despite the initial difficulties (Corpas Pastor 2008: 27).

In the 1990s, thanks to technological developments, researchers were able to compile large reference corpora. One such example is the pioneering *British National Corpus* (1995), which includes 100 million words. At this time, corpora began to be used beyond lexicography and linguistics, such as in translation studies. This period saw the arrival of "mega-corpora" (Zanettin 2012: 8). Indeed, huge corpora have been created in the last few years and, fortunately, many of them are now accessible via the Internet.

The number of Basque corpora has been growing continuously since the beginning of the twenty-first century. Consequently, the Basque (research) community now has access to high-quality Basque reference corpora. This will be demonstrated in the following section. However, there was no corpus of literary texts translated from German into Basque. Given that the aim of my research is the analysis of these types of texts, it was necessary to build a digitised, parallel and multilingual corpus from scratch. Section 3 of this chapter describes the process of designing, compiling and annotating the corpus. Initially, the corpus did not include any form of lemmatization or part-of-speech (POS) tagging. This was not necessary when examining the type of phraseological units under analysis. However, because the goal is now to extract Basque multi-word expressions (MWEs) with a particular structure (see Section 5), the corpus has been lemmatized and tagged at POS level (as shown in Section 4). Following the extraction and selection of the MWEs, they will then be analysed from a translational perspective.

2. An overview of Basque corpora

The first corpus – the *OEH* text corpus (see Table 1) – was constructed in 1984. It seems that Basque was not far behind other languages in the creation of corpora (Areta et al. 2008: 79). Nevertheless, there was a gap of 18 years before the creation of the next corpus, the Statistical Corpus of the Twentieth Century. Basque corpora consisting of a million words arrived later than in other languages. Regardless of the time lapse, the first *modern* Basque corpus, the Statistical Corpus of the Twentieth Century, was, and still is, of great value. It served as a model for subsequent corpora. Small as it was from a contemporary perspective, it remains the only Basque corpus representing the language of the twentieth century (Urkia 2010: 6).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 1. List of representative Basque corpora

	Corpus	Date	Author	Size (million words)	Language
1	Corpus of the *Orotariko Euskal Hiztegia* (OEH)	1984	The Royal Academy of the Basque Language (*Euskaltzaindia*)	5.6	Basque
2	The Statistical Corpus of the Twentieth Century	2002	The Royal Academy of the Basque Language (*Euskaltzaindia*)	5	Basque
3	LEGE-bi	2004	University of Deusto	2.4	Basque, Spanish
4	The Corpus of Classical Works (*Klasikoen gordailua*)	2005	Susa	11.9	Basque
5	The Science and Technology Corpus (*Zientzia eta Teknologia corpusa*)	2006	Group IXA and Elhuyar	8.5	Basque
6	The Corpus of Reference and Contemporary Prose (*Ereduzko Prosa Gaur*)	2007	Ibon Sarasola, Pello Salaburu, Josu Landa and Josu Zabaleta	25.1	Basque
7	The Dynamic Corpus of Reference and Contemporary Prose (*Ereduzko Prosa Dinamikoa*)	2009	Ibon Sarasola, Pello Salaburu and Josu Landa	Around 15	Basque
8	Science for all Readers (*Zientzia Irakurle Ororentzat*)	2010	Ibon Sarasola, Pello Salaburu and Josu Landa	1.3	Basque
9	Eroski Consumer	2010	Elhuyar and Eleka	15.4	Basque, Spanish, Catalan and Galician
10	Lexical Observatory (*Lexikoarean Behatokiaren Corpusa*)	2010	The Royal Academy of the Basque Language (*Euskaltzaindia*)	50.3	Basque
11	*Goenkale corpusa* (a Basque soap opera)	2011	Ibon Sarasola, Pello Salaburu and Josu Landa	11	Basque
12	Classical Works (*Pentsamenduaren klasikoak corpusa*)	2011	Ibon Sarasola, Pello Salaburu and Josu Landa	10.7	Basque
13	Translational Database of the Gipuzkoan Provincial Government	2011	Gipuzkoan Provincial Government	Constantly updated	Basque, Spanish

(*continued*)

Table 1. (*continued*)

	Corpus	Date	Author	Size (million words)	Language
14	*Bizimena* (Translational Database of the Biscayan Provincial Government)	2011	Biscayan Provincial Government	Constantly updated	Basque, Spanish
15	Corpus of Legal Texts (*Zuzenbide corpusa*)	2012	Joseba Ezeiza	7.2	Basque
16	EHUskaratuak (Corpus of Translated Textbooks)	2012	Elhuyar	18	Basque, Spanish, English, French
17	Bilingual Web Corpus (*Web Corpusen Ataria*)	2013	Elhuyar	18.8	Basque, Spanish
18	Monolingual Web Corpus (*Web Corpusen Ataria*)	2013	Elhuyar	125	Basque
19	Corpus of Contemporary Basque (*Egungo Testuen Corpusa*)	2013	Ibon Sarasola, Pello Salaburu and Josu Landa	269.2	Basque
20	Corpus of (mostly) Translated Texts (*Hizkuntzen arteko corpusa*)	2015	Ibon Sarasola, Pello Salaburu and Josu Landa	42.43	Basque, Spanish, French, English

Table 1 shows the most representative corpora (as far as size is concerned) of written Basque, presented chronologically. The corpora created between 1984 and 2010 were extracted from Areta et al. (2008) and Urkia (2010). The rest were added by the author of this chapter.

Of the 20 corpora that are included in Table 1, seven are bi- or multilingual. The first is the specialized LEGE-bi corpus,[1] which was created at the University of Deusto in 2004. Chronologically, the next multilingual corpus is the Eroski Consumer Corpus.[2] This is a parallel corpus consisting of texts in four different languages: Basque, Spanish, Catalan and Galician. This corpus represents remarkable progress in the creation of Basque multilingual, parallel corpora. Two specialised translational databases from the corresponding provincial governments

[1] <http://sli.uvigo.es/CLUVI/corpus_en.html>

[2] <http://corpus.consumer.es/corpus/>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

in Gipuzkoa[3] and Biscay[4] have been accessible since 2011. These contain legal and administrative texts. EHUskaratuak is another noteworthy corpus, which was launched in 2002.[5] It consists of English, French and Spanish academic source texts translated into Basque from 2007 onwards. The Elhuyar foundation published the website *Web Corpusen Ataria,*[6] the following year (in 2013). The website includes three independent applications: a Spanish and Basque parallel corpus, a monolingual Basque corpus and an application for the automatic extraction of Basque bigrams. Both corpora are innovative in that they comprise texts that are automatically extracted from the Internet. According to a member of the foundation (Leturia Azkarate 2013), this method of creating corpora has enabled them to build large corpora in an easy, fast and economic way. Although they admit that the corpus may include poor-quality texts, they argue that this reflects the real use of language. According to them, this is the ultimate purpose of using corpora. The latest multilingual corpus is the Corpus of Translated Texts. This mainly comprises translated texts in Basque, Spanish, French, and English, alongside some original texts.

Table 1 also includes quite a few specialized corpora. Some contain scientific texts, such as the Science and Technology Corpus[7] and Science for all Readers.[8] The Corpus of Legal Texts contains judicial texts.[9] Another two corpora are made up of classical texts,[10] and the *Goenkale corpusa*[11] consists of dialogues from a popular Basque television soap opera. This corpus (as stated on its website) offers a unique opportunity to analyse how Basque behaves in another register, namely dialogue.

The most significant corpora with regard to size will be mentioned next. The Corpus of Reference and Contemporary Prose[12] comprises two subcorpora: one consists of literary texts (2000–2006) and the other, of press texts (2001–2006).

[3] <http://www.gipuzkoa.eus/imemoriak/>

[4] <apli.bizkaia.net/Apps/Danok/CEMT/Default.aspx?locale=es-ES>

[5] <http://ehuskaratuak.ehu.eus/kontsulta/>

[6] <http://webcorpusak.elhuyar.eus/>

[7] <http://www.ztcorpusa.eus/cgi-bin/kontsulta.py>

[8] <http://www.ehu.eus/ehg/zio/>

[9] <http://www.ehu.eus/ehg/zuzenbidea/>

[10] <http://www.ehu.eus/ehg/pkc/ and http://www.ehu.eus/ehg/kc/>

[11] <http://www.ehu.eus/ehg/goenkale/>

[12] <http://www.ehu.eus/euskara-orria/euskara/ereduzkoa/>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

An advantage of this corpus is the similar size of both subcorpora. This enables the user to easily compare their results.[13]

Aside from web corpora, the largest Basque corpora are the Lexical Observatory[14] and the Corpus of Contemporary Basque.[15] The Lexical Observatory is a monitor corpus consisting of press texts. It was created at the initiative of *Euskaltzaindia*, the Royal Academy of the Basque Language. As explained on their website, the goal is to create a large, balanced, lemmatized and linguistically annotated corpus, although they have not yet achieved this aim. The Corpus of Contemporary Basque is currently the largest existing Basque reference corpus. It is a balanced corpus in terms of text type, date of publication and translated/non-translated texts. However, there are remarkable differences in the amount of words for the different text types. For instance, texts labelled as press texts make up 115.5 million words whereas literary texts are made up of 35.6 million words.

3. Design, compilation and annotation of the Aleuska corpus

Prior to the compilation of my own corpus, it was necessary to create a catalogue of all literary works that have been translated from German into Basque. The Aleuska catalogue is updated to the year 2013 and contains 710 entries. These are divided into different text types: children's literature, adult literature, essay, poetry, and theatre.

The Aleuska catalogue was described in detail. Following this, the results (partly explained in the following) served to establish the criteria for selecting texts to be included in the corpus. The description shows, for instance, that translation activity began to be significant (concerning the quantity of works translated from German into Basque) from the year 1980 onwards. Thus, texts translated from the 1980s on were included in the corpus. Secondly, an attempt was made at the catalogue level to identify the mode of translation. Had the texts been directly translated from the German source text, or were they indirect translations from an intermediary text (a Spanish version in most cases)? Although the results obtained from analyses of paratextual information regarding this feature were

[13] The Dynamic Corpus of Reference and Contemporary Prose <http://www.ehu.eus/ehg/epd/> was created as a complement of the Corpus of Reference and Contemporary Prose. While the latter is a static corpus, the former is a dynamic corpus that is updated yearly because it aims to reflect contemporary language use.

[14] <http://lexikoarenbehatokia.euskaltzaindia.eus/cgi-bin/kontsulta.py>

[15] <http://www.ehu.eus/etc/>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

inconclusive, some tendencies were observed. It was decided that both direct and indirect translations would be included in the corpus.[16] Thirdly, both adult and children's literary texts (representing 71% of the catalogue), were included. Finally, it was important to ensure that the corpus was diverse in terms of source and target authors. As such, texts were included from a wide variety of original authors and translators.

The specific features of the corpus will be shown later in this chapter. For now, it is sufficient to mention that 24 adult literary texts (AL) and 24 children's literary texts (CL) were selected. For the purpose of compiling the corpus, these texts were first digitised and cleaned. Once we had the cleaned TXT files for each text, they were tagged and aligned with a program called TAligner. This has been developed (and is still under further development) within the Tralima-Itzulik research group[17] at the University of the Basque Country. TAligner is a tool written in Java and developed by Iñaki Albisua.

We decided to work with our own tool and not with other tools commercially or non-commercially available. This was mainly due to a specific feature of our corpus: up to three texts had to be simultaneously aligned. It was possible to adapt TAligner according to our specific needs.[18]

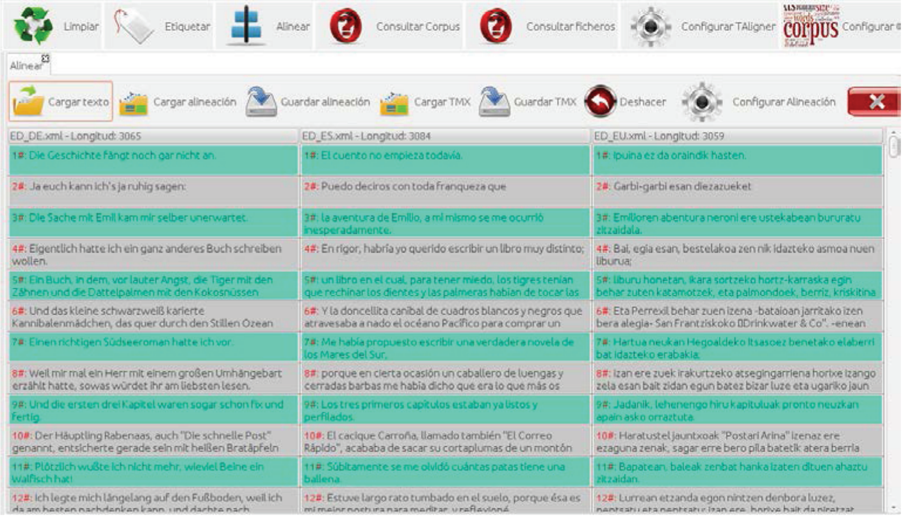
Figure 1 shows how three aligned texts appear in TAligner: the source German text, the intermediary Spanish text, and the Basque translation. In this example, the sentences have already been adjusted. This means that the program user needs to make manual adjustments to correctly align the texts: divide cells, merge cells, add cells etc. Although time-consuming, this allows us to minimise alignment problems when querying the corpus and obtaining results.

[16] It is not the aim of the present chapter to give a detailed description of the Aleuska catalogue. However, regarding direct and indirect translations, it is noteworthy that indirect translations are a reality in German-into-Basque translation (predominantly in children's literaure), and therefore, they needed to be included in the corpus.

[17] <http://www.ehu.eus/tralima/inicio_eng.php>

[18] Serón Ordoñez (2015) describes another procedure to create corpora consisting of original texts and more than one target text. The advantages of TAligner are that the user can align at once as many target texts as he/she wishes, and alignment and queries can be conducted using this one tool.

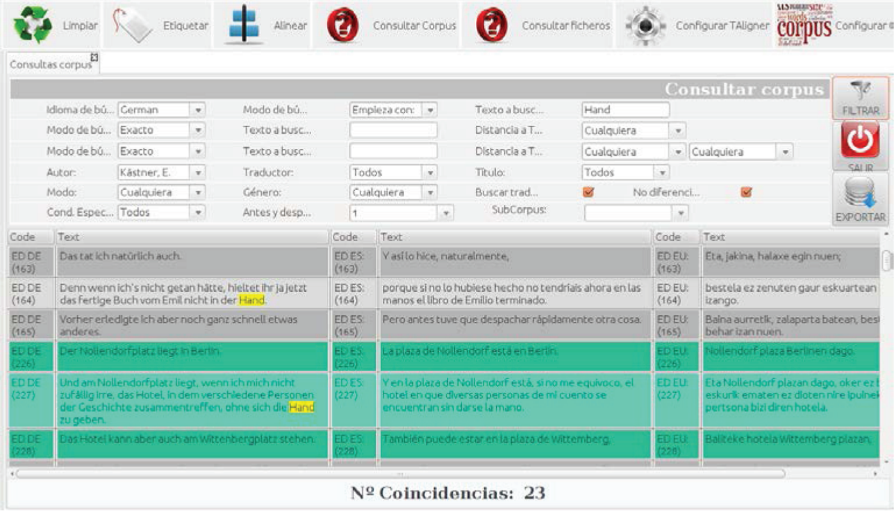
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



ED_DE.xml - Longitud: 3065	ED_ES.xml - Longitud: 3084	ED_EU.xml - Longitud: 3059
1#: Die Geschichte fängt noch gar nicht an.	1#: El cuento no empieza todavía.	1#: Ipuina ez da oraindik hasten.
2#: Ja euch kann ich's ja ruhig sagen:	2#: Puedo deciros con toda franqueza que	2#: Garbi-garbi esan diezazueket
3#: Die Sache mit Emil kam mir selber unerwartet.	3#: la aventura de Emilio, a mí mismo se me ocurrió inesperadamente.	3#: Emilioren abentura neroni ere ustekabean bururatu zitzaidala.
4#: Eigentlich hatte ich ein ganz anderes Buch schreiben wollen.	4#: En rigor, habría yo querido escribir un libro muy distinto;	4#: Bai, egia esan, bestelakoa zen nik idazteko asmoa nuen liburua;
5#: Ein Buch, in dem, vor lauter Angst, die Tiger mit den Zähnen und die Dattelpalmen mit den Kokosnüssen	5#: un libro en el cual, para tener miedo, los tigres tenían que rechinar los dientes y las palmeras habían de tocar las	5#: liburu honetan, ikara sortzeko hortz-karraska egin behar zuten katamotzek, eta palmondoek, berriz, kriskitina
6#: Und das kleine schwarzweiß karierte Kannibalenmädchen, das quer durch den Stillen Ozean	6#: Y la doncellita caníbal de cuadros blancos y negros que atravesaba a nado el océano Pacífico para comprar un	6#: Eta Perrexil behar zuen izena -bataioan jarritako izen bera alegia- San Frantziskoko ▯Drinkwater & Co". -enean
7#: Einen richtigen Südseeroman hatte ich vor.	7#: Me había propuesto escribir una verdadera novela de los Mares del Sur.	7#: Hartua neukan Hegoaldeko Itsasoez benetako eleberri bat idazteko erabakia.
8#: Weil mir mal ein Herr mit einem großen Umhängebart erzählt hatte, sowas würdet ihr am liebsten lesen.	8#: porque en cierta ocasión un caballero de luengas y cerradas barbas me había dicho que era lo que más os	8#: Izan ere zuek irakurtzeko atsegingarriena horixe izango zela esan bait zidan egun batez bizar luze eta ugariko jaun
9#: Und die ersten drei Kapitel waren sogar schon fix und fertig.	9#: Los tres primeros capítulos estaban ya listos y perfilados.	9#: Jadanik, lehenengo hiru kapituluak pronto neuzkan apain asko orraztuta.
10#: Der Häuptling Rabenaas, auch "Die schnelle Post" genannt, entsicherte gerade sein mit heißen Bratäpfeln	10#: El cacique Carroña, llamado también "El Correo Rápido", acababa de sacar su cortaplumas de un montón	10#: Haratustel jauntxoak "Postari Arina" izenaz ere ezaguna zenak, sagar erre bero pila batetik atera berria
11#: Plötzlich wußte ich nicht mehr, wieviel Beine ein Walfisch hat!	11#: Súbitamente se me olvidó cuántas patas tiene una ballena.	11#: Bapatean, baleak zenbat hanka izaten dituen ahaztu zitzaidan.
12#: Ich legte mich längelang auf den Fußboden, weil ich	12#: Estuve largo rato tumbado en el suelo, porque ésa es	12#: Lurrean etzanda egon nintzen denbora luzez,

Figure 1. The alignment of three texts in TAligner

Figure 2 shows the search engine, indicating how queries are made in the corpus. In this case the query was limited to German words that began with *Hand* in Erich Kästner's books. The result shows the German sentence that contains the word *Hand*, together with the previous and subsequent sentences, as well as the Spanish and Basque counterparts. In this case, the Spanish version was included because this text was translated indirectly.



Idioma de bú...	German	Modo de bú...	Empieza con:	Texto a busc...	Hand	
Modo de bú...	Exacto	Texto a busc...		Distancia a T...	Cualquiera	
Modo de bú...	Exacto	Texto a busc...		Distancia a T...	Cualquiera	Cualquiera
Autor:	Kästner, E.	Traductor:	Todos	Título:	Todos	
Modo:	Cualquiera	Género:	Cualquiera	Buscar trad...	☑	No diferenci... ☑
Cond. Espec...	Todos	Antes y desp...	1	SubCorpus:		

Code	Text	Code	Text	Code	Text
ED DE (163)	Das tat ich natürlich auch.	ED ES: (163)	Y así lo hice, naturalmente,	ED EU: (163)	Eta, jakina, halaxe egin nuen;
ED DE (164)	Denn wenn ich's nicht getan hätte, hieltet ihr ja jetzt das fertige Buch vom Emil nicht in der Hand.	ED ES: (164)	porque si no lo hubiese hecho no tendríais ahora en las manos el libro de Emilio terminado.	ED EU: (164)	bestela ez zenuten gaur eskuartean izango.
ED DE (165)	Vorher erledigte ich aber noch ganz schnell etwas anderes.	ED ES: (165)	Pero antes tuve que despachar rápidamente otra cosa.	ED EU: (165)	Baina aurretik, zalaparta batean, besi behar izan nuen.
ED DE (226)	Der Nollendorfplatz liegt in Berlin.	ED ES: (226)	La plaza de Nollendorf está en Berlín.	ED EU: (226)	Nollendorf plaza Berlinen dago.
ED DE (227)	Und am Nollendorfplatz liegt, wenn ich mich nicht zufällig irre, das Hotel, in dem verschiedene Personen der Geschichte zusammentreffen, ohne sich die Hand zu geben.	ED ES: (227)	Y en la plaza de Nollendorf está, si no me equivoco, el hotel en que diversas personas de mi cuento se encuentran sin darse la mano.	ED EU: (227)	Eta Nollendorf plazan dago, oker ez t eskurik ematen ez dioten nire ipuinek pertsona bizi diren hotela.
ED DE (228)	Das Hotel kann aber auch am Wittenbergplatz stehen.	ED ES: (228)	También puede estar en la plaza de Wittemberg.	ED EU: (228)	Baliteke hotela Wittemberg plazan.

Nº Coincidencias: 23

Figure 2. Querying the corpus in the TAligner program

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 2 summarizes features of the corpus. As was mentioned previously, 24 AL texts and 24 CL texts were included in the corpus. The number of direct translations is higher because the main interest of this study lies in the results obtained from direct translations. However, indirect translations are also a reality and, because it is a highly interesting phenomenon to observe, 14 indirect translations were included. In addition, the corpus is made up of works of 30 different German original authors and 28 different Basque translators. The last three columns represent the number of words per language. The Spanish translations are only included when it is assumed that the texts were translated indirectly,[19] which explains the low number. All in all, the corpus contains around 3.5 million words. Although it is a fairly small corpus, it is large enough to reach the objective of the study.

Table 2. Overview of the Aleuska corpus

	Mode of translation		**Authors**		**Number of words**		
	Direct	Indirect	Source authors	Target authors	German	Basque	Spanish
AL	19	5	17	15	1,120,534	935,530	198,274
CL	15	9	13	16	593,871	512,204	166,860
Total	34	14	30	28	1,714,405	1,447,734	365,134

After the completion of my thesis, one of the goals was to lemmatize and POS tag the corpus. Up until then, annotation was limited to the segmentation of the texts into sentences and paragraphs. IXA pipe tools were used to lemmatize and POS tag the corpus. This allows researchers the opportunity to linguistically annotate corpora without having detailed knowledge of NLP (Agerri et al. 2014: 3823).

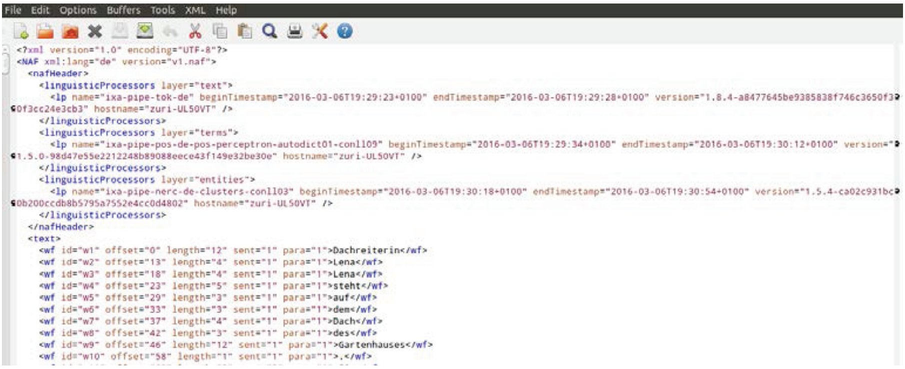
The following annotation options are offered in different languages:

- Tokenization: German, Basque, French, Galician, Spanish, English, Italian and Danish.
- Lemmatization and POS tagging: German, Basque, French, Galician, Spanish, English, Italian and Danish.
- Named Entity Recognition: German, Basque, Spanish, English, Italian and Danish.
- Chunking: Basque and English.
- Parser: Spanish and English.

[19] When creating the catalogue, publishers and translators were contacted to obtain information regarding the mode of translation; that is, when possible, they were asked about the source version(s) from which the Basque translations were made.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

In the case of my corpus, these tools were used to lemmatize and tag the German, Basque and Spanish texts. The steps that are necessary to conduct all three processes are well explained on the IXA pipes website <http://ixa2.si.ehu.es/ixa-pipes/>. Input files are always raw TXT files and a lemmatized NAF tagged file will be obtained for each text. If there is a large number of texts to be tagged, then it is advisable to implement a simple script for each language. With this script, all texts in a certain language can be tokenized, lemmatized and POS tagged at the same time. This is possible thanks to the Unix pipes metaphor, according to which the output of each process is fed directly as input for the next <http://ixa2.si.ehu.es/ixa-pipes/>. A sample of how tokenization has been carried out for a German text can be seen at the top of the sample NAF file displayed in Figure 3.



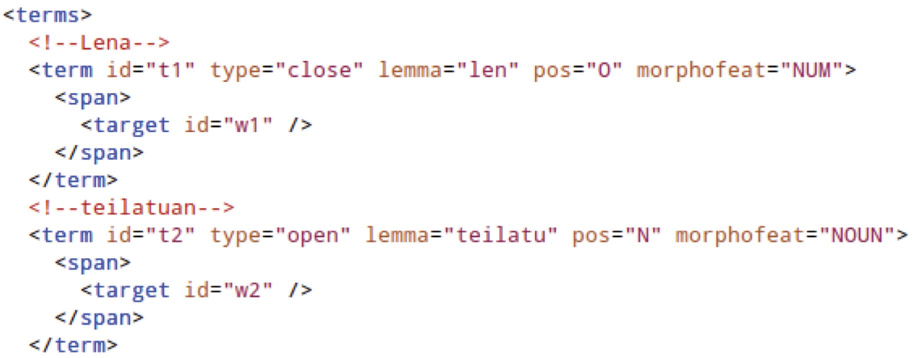
```
<?xml version="1.0" encoding="UTF-8"?>
<NAF xml:lang="de" version="v1.naf">
  <nafHeader>
    <linguisticProcessors layer="text">
      <lp name="ixa-pipe-tok-de" beginTimestamp="2016-03-06T19:29:23+0100" endTimestamp="2016-03-06T19:29:28+0100" version="1.8.4-a8477645be9385838f746c3650f3
0f3cc24e3cb3" hostname="zuri-UL50VT" />
    </linguisticProcessors>
    <linguisticProcessors layer="terms">
      <lp name="ixa-pipe-pos-de-pos-perceptron-autodict01-conll09" beginTimestamp="2016-03-06T19:29:34+0100" endTimestamp="2016-03-06T19:30:12+0100" version="
1.5.0-98d47e55e2212248b89088eece43f149e32be30e" hostname="zuri-UL50VT" />
    </linguisticProcessors>
    <linguisticProcessors layer="entities">
      <lp name="ixa-pipe-nerc-de-clusters-conll03" beginTimestamp="2016-03-06T19:30:18+0100" endTimestamp="2016-03-06T19:30:54+0100" version="1.5.4-ca02c931bc
0b200ccdb8b5795a7552e4cc0d4802" hostname="zuri-UL50VT" />
    </linguisticProcessors>
  </nafHeader>
  <text>
    <wf id="w1" offset="0" length="12" sent="1" para="1">Dachreiterin</wf>
    <wf id="w2" offset="13" length="4" sent="1" para="1">Lena</wf>
    <wf id="w3" offset="18" length="4" sent="1" para="1">Lena</wf>
    <wf id="w4" offset="23" length="5" sent="1" para="1">steht</wf>
    <wf id="w5" offset="29" length="3" sent="1" para="1">auf</wf>
    <wf id="w6" offset="33" length="3" sent="1" para="1">dem</wf>
    <wf id="w7" offset="37" length="4" sent="1" para="1">Dach</wf>
    <wf id="w8" offset="42" length="3" sent="1" para="1">des</wf>
    <wf id="w9" offset="46" length="12" sent="1" para="1">Gartenhauses</wf>
    <wf id="w10" offset="58" length="1" sent="1" para="1">.</wf>
```

Figure 3. An excerpt of a NAF file showing the tokenization

The tokens corresponding to the text are stored between the tags <*text*> . Several pieces of information are saved for each of the tokens: an id number to identify the token, the length of the token, as well as the sentence and paragraph numbers in which the token appears. The information regarding the part of speech is saved between the opening and closing tags <*terms*> , as can be observed in the Basque target text in Figure 4.

In this case, information is provided for each term regarding its id, type, lemma, the current POS tag, and the morphological feature. This simple example shows that the automatic tagging is not error free. For instance, the lemma of Lena is identified as “len” (instead of “lena”) and it is identified as a numeral (NUM) and not as a proper noun (PROPN). A review of the output obtained from automatic tagging will be necessary in the future. Once the texts have been tagged at POS level, the next step is to extract the specific MWEs. The next section will describe this process.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



```
<terms>
  <!--Lena-->
  <term id="t1" type="close" lemma="len" pos="O" morphofeat="NUM">
    <span>
      <target id="w1" />
    </span>
  </term>
  <!--teilatuan-->
  <term id="t2" type="open" lemma="teilatu" pos="N" morphofeat="NOUN">
    <span>
      <target id="w2" />
    </span>
  </term>
```

Figure 4. An excerpt of a NAF file showing the POS tagging

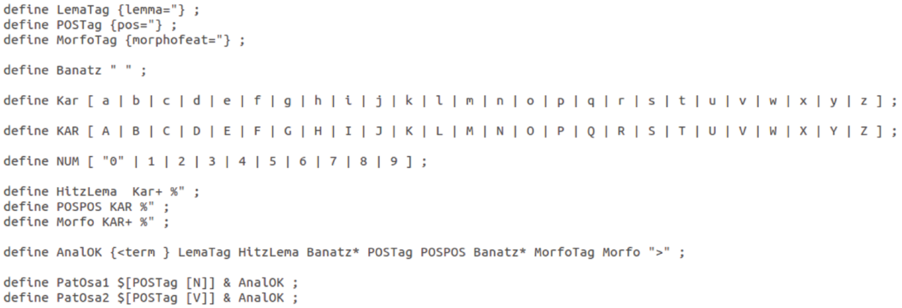
4. Extraction of MWEs

My PhD thesis analysed the German-into-Basque translation of a specific type of MWE, that is, somatic phraseological units and binomials. The goal now is to expand the analysis to certain Basque MWEs that are very frequent in both common and educated language (Altzibar et al. 2011). Here I refer to collocations formed by an onomatopoeic expression and a verb. For example, *zanga-zanga edan*, which according to the Basque onomatopoeia dictionary (Ibarretxe Antuñano and Martinez Lizarduikoa 2006) means *slurping down* or *drinking in gulps*. The primary goal will be to determine how and when these common Basque collocations are used in the present corpus and which differences and similarities can be established between direct and indirect translations, adult and children's literature, and between translated and non-translated language.[20]

The toolkit used to extract the patterns is called Foma (Hulden, 2009). This is a free and open-source tool created and maintained by Mans Hulden which may have a variety of uses within natural language processing. With the assistance of Iñaki Alegria (a member of the research group IXA at the University of the Basque Country), Foma was used to write and process code to extract user-defined patterns (based on information provided by POS tags in NAF files). Figure 5 shows an excerpt of this code.

[20] This last feature – determining differences and similarities between translated and non-translated language – is regarded as a long-term goal because we still do not have our own Basque monolingual corpus of literary texts. However, the existing corpora that were mentioned in the first section would be of great help to carry out this analysis.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



```
define LemaTag {lemma="} ;
define POSTag {pos="} ;
define MorfoTag {morphofeat="} ;

define Banatz " " ;

define Kar [ a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z ] ;

define KAR [ A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z ] ;

define NUM [ "0" | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 ] ;

define HitzLema  Kar+ %" ;
define POSPOS KAR %" ;
define Morfo KAR+ %" ;

define AnalOK {<term } LemaTag HitzLema Banatz* POSTag POSPOS Banatz* MorfoTag Morfo ">" ;

define PatOsa1 $[POSTag [N]] & AnalOK ;
define PatOsa2 $[POSTag [V]] & AnalOK ;
```

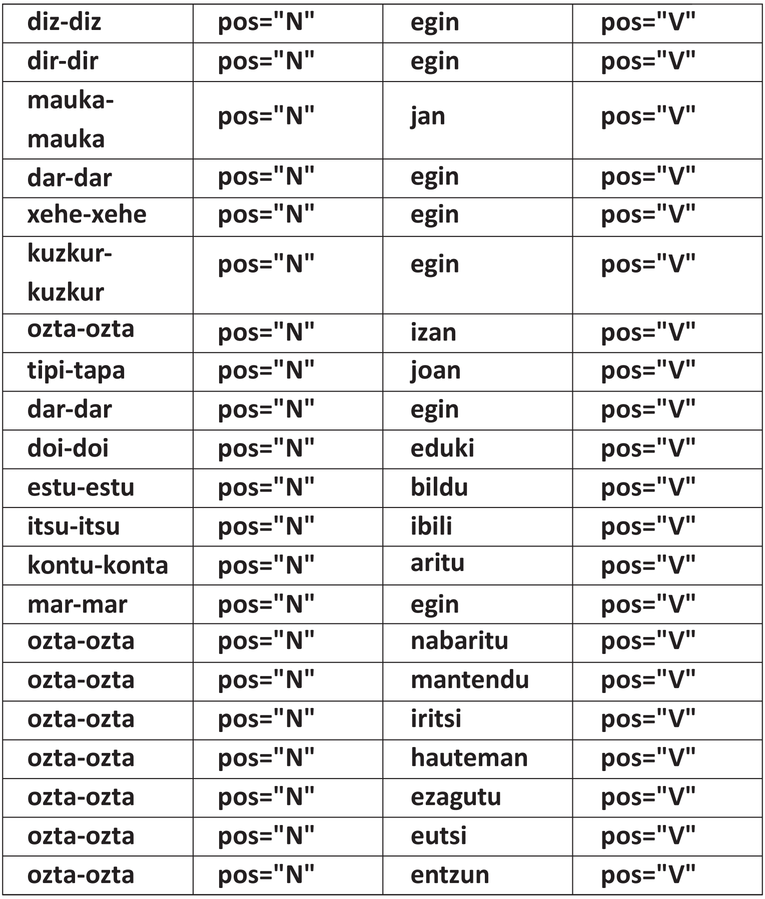
Figure 5. An excerpt of the Foma code

Firstly, tag information was defined as mentioned in the previous section: *lemma* (*LemaTag*), *pos* (*POSTag*) and *morphofeat* (*MorfoTag*). The description of spaces (*Banatz*) and characters (*Kar*, *KAR*, *NUM*) is necessary to specify the content of lemmas (*HitzLema*), POS tags (*POSPOS*) and morphological features (*Morfo*). The *AnalOK* rule contains the line of the NAF file where lemmatization and POS tagging occurs: the tag <*term*, the lemma tag, the lemma itself, space, the POS tag, the content of the POS tag, space, the morphofeat tag, its content and the closing tag > . The following two rules of the code serve to define the patterns to be extracted. There are several more rules in the code that are used to let Foma know how results should be presented. After compiling the Foma file with this code, we will be able to extract the patterns defined in it with the help of a script.

As expected, the number of extracted patterns was very large. The list was shortened considerably by only selecting patterns that included a hyphen. This method was chosen as the type of MWE being studied (onomatopoeia+verb) always contains a hyphen in the onomatopoeic part. From the 2,617 patterns containing a hyphen, I manually selected the potential collocations (totalling 147 different types).

Figure 6 illustrates some of the extracted MWEs containing an onomatopoeic expression and a verb in Basque. For example, *diz-diz egin* (to sparkle), *dir-dir egin* (to sparkle), *mauka-mauka jan* (to eat voraciously) or *dar-dar egin* (to tremble). The next step is to make queries in the TAligner program containing these MWEs to obtain the German (and Spanish in the case of indirect translations) correspondences. One could then draw some conclusions as regards the differences and similarities between direct and indirect translations, adult and children's literature, and translations and non-translations.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use



diz-diz	pos="N"	egin	pos="V"
dir-dir	pos="N"	egin	pos="V"
mauka-mauka	pos="N"	jan	pos="V"
dar-dar	pos="N"	egin	pos="V"
xehe-xehe	pos="N"	egin	pos="V"
kuzkur-kuzkur	pos="N"	egin	pos="V"
ozta-ozta	pos="N"	izan	pos="V"
tipi-tapa	pos="N"	joan	pos="V"
dar-dar	pos="N"	egin	pos="V"
doi-doi	pos="N"	eduki	pos="V"
estu-estu	pos="N"	bildu	pos="V"
itsu-itsu	pos="N"	ibili	pos="V"
kontu-konta	pos="N"	aritu	pos="V"
mar-mar	pos="N"	egin	pos="V"
ozta-ozta	pos="N"	nabaritu	pos="V"
ozta-ozta	pos="N"	mantendu	pos="V"
ozta-ozta	pos="N"	iritsi	pos="V"
ozta-ozta	pos="N"	hauteman	pos="V"
ozta-ozta	pos="N"	ezagutu	pos="V"
ozta-ozta	pos="N"	eutsi	pos="V"
ozta-ozta	pos="N"	entzun	pos="V"

Figure 6. Potential Basque MWE consisting of an onomatopoeia and a verb

5. Conclusion

As discussed at the beginning of this chapter, the first Basque corpus was compiled comparatively early. With regard to today's standards it is quite small. Up to 2010, Basque corpora were not generally very large. However, it is important to bear in mind that we are dealing with a minority language. Despite this limitation, a wide range of very diverse (monolingual, multilingual, general, specialised) corpora have been presented here. This can be of benefit to anyone wanting to work with Basque corpora.

Since German is not included in any of the corpora in Table 1, I needed to create my own parallel and multilingual corpus consisting of literary texts translated (both directly and indirectly) from German into Basque. In order to achieve this, I

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

used a methodology starting with the compilation of a catalogue to the creation of the corpus itself. It is important to mention that the TAligner program was essential to build the multilingual and parallel corpus. This is due to its user-friendliness and flexibility concerning the amount of text that can be simultaneously aligned. The resulting corpus includes literary texts from adult and children's literature, with direct and also indirect translations from intermediary texts. This ensures diversity in terms of source and target authors as well as the publication year of the Basque translations. In terms of size, this is not a large corpus but it is large enough for my research purposes.

For the lemmatization, POS tagging, and extraction of MWEs, I used free and open-source resources, such as the IXA pipe tools and Foma (created by Mans Hulden). A future goal is to extract sentences in which the usual word combinations are embedded in the Basque translations together with their equivalents in the source text(s), then to analyse in which cases these potential MWEs were used in the texts. For now, the extraction of MWEs was limited to Basque collocations containing an onomatopoeic expression and a verb due to their singularity and their frequency in Basque common and educated language.

To conclude, a variety of tools were required to carry out this corpus-based translation analysis of MWEs. One of the future goals of this research is to *merge* all of these tools and to have an all-in-one tool that will enable the researcher to clean and align the corpus, make queries in it, lemmatize and tag it at different levels, and extract the desired patterns from it.

References

Altzibar, Xabier & Bilbao, Xabier & Garai, Koldo. 2011. Collocations in Basque: A test for classification. In *Proceedings of the 5th International Conference on Meaning-Text Theory*, Barcelona, September 8–9, 1–12.

Agerri, Rodrigo & Bermudez, Josu & Rigau, German. 2014. IXA pipeline: efficient and ready to use multilingual NLP Tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, May 26–31.

Areta, Nerea & Gurrutxaga, Anton & Leturia, Igor. 2008. Begiratu bat corpus-baliabideei. *BAT Soziolinguistika aldizkaria* 62: 71–92.

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

Hulden, Mans. 2009. Foma: A finite-state toolkit and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 29–32.

Ibarretxe Antuñano, Iraide & Martinez Lizarduikoa, Alfontso. 2006. *Hizkuntzaren bihotzean: Euskal onomatopeien hiztegia*. Donostia-San Sebastian: Gaiak.

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation. A corpus-based approach*. Manchester: St. Jerome.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Leturia Azkarate, Igor. 2013. Web-corpusen Ataria. *Elhuyar aldizkaria* 13(03): 294–295.

Serón Ordóñez, Inmaculada. 2015. Cómo crear y analizar corpus paralelos. Un procedimiento con *software* accesible y económico y algunas sugerencias para *software* futuro. In *Corpus-based Translation and Interpreting Studies*: *From description to application*, María Teresa Sánchez Nieto (ed). Berlin: Frank & Timme. 167–190.

Sinclair, John. 2005. Corpus and text-basic principles. In *Developing Linguistic Corpora: A Guide to Good Practice*, Martin Wyne (ed). Oxford: University of Oxford–AHDS Literature, Languages and Linguistics. <http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm> (6 May 2017).

Urkia, Miriam. 2010. Corpusgintzaren garrantzia hizkuntzalaritzan eta euskararen egoeran. <http://www.euskaltzaindia.eus/dok/plazaberri/2010/urtarrila/corpusgintza_miriamurkia.pdf> (6 May 2017).

Zanettin, Federico. 2012. *Translation-driven Corpora*. Manchester: St. Jerome.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART III

Parallel corpora

Tools and applications

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Strategies for building high quality bilingual lexicons from comparable corpora

Pablo Gamallo
University of Santiago de Compostela

This chapter outlines two strategies to automatically build bilingual dictionaries: One is based on the use of a pivot language and existing bilingual dictionaries, while the other relies on string similarity and cognate extraction. Both strategies have in common the use of translation equivalents extracted from comparable corpora to filter out odd bilingual pairs and validate the correct ones. The correctness of the entries validated with comparable corpora is very high, close to that achieved by using parallel corpora. The chapter reports several case studies describing how to build new high-quality bilingual lexicons, namely English-Galician, English-Portuguese, and Portuguese-Spanish dictionaries with more than 90% precision. This outperforms state-of-the-art systems on bilingual extraction from comparable corpora, whose best scores hardly reach 70 or 80%.

Keywords: comparable corpora, extraction of translation candidates, bilingual lexicons, distributional similarity, cognates

1. Introduction

A comparable corpus consists of documents in two or more languages or varieties which are not a translation of each other and deal with similar topics. Comparable corpora are by definition multilingual and cross-lingual text collections. The use of comparable corpora to automatically extract bilingual lexicons has been growing in recent years (Tamura et al. 2012; Aker et al. 2013; Ansari et al. 2014; Hazem & Morin 2014). The main advantage of using comparable corpora to perform this extraction task is that they are easily available and make use of the internet as a huge resource of multilingual texts. Comparable corpora are more easily available than parallel texts, especially for minority languages. However, their main drawback is the low performance of the extraction systems based on them. According to Nakagawa (2001), bilingual lexicon extraction from comparable corpora is an

https://doi.org/10.1075/scl.90.15gam

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

© 2019 John Benjamins Publishing Company

overly difficult and ambitious objective, and much more complex than extraction from parallel and aligned corpora.

It is possible, however, to use comparable corpora in a less ambitious way, not to build large and accurate bilingual lexicons from scratch, but just to filter out false bilingual pairs from those selected by other basic bilingual extraction methods. In this chapter, we will focus on two basic methods for extracting bilingual lexicons: first, the construction of new bilingual dictionaries by transitivity using intermediary dictionaries and, second, the selection of bilingual cognates by means of string similarity. These two strategies are aimed at building large bilingual lexicons and/or terminologies, even if their correctness and precision is low due to polysemy and false friend candidates. In order to discard false pairs and select only correct pairs of translation candidates, we take into account their context and distribution in comparable corpora. In this way, the bilingual pairs with a similar distribution in comparable corpora are considered to be correct pairs and subsequently are not removed from the lexicon.

Let us see some examples to illustrate our methodology. Suppose we are deriving a new English-Galician dictionary by crossing two existing ones, English-Spanish and Spanish-Galician, with Spanish as the intermediary (or pivot) language between English and Galician. This is a transitive operation which consists of generating bilingual correspondences for a new language pair from two already known language pairs. Let us take, for instance the English verb *subside*, which is translated by the polysemous Spanish word *bajar* in the English-Spanish dictionary. The Spanish polysemous word is, in turn, translated in the Spanish-Galician dictionary into two different verbs: *baixar* (*go down*) and *apear* (*take down*). Then, the derived English-Galician dictionary generates the bilingual pairs (*subside, baixar*) and (*subside, apear*). While the former translation is correct, the latter is clearly odd. The Galician verb *apear* does not mean *subside* in any context; it means *take down*, which is one of the senses of the Spanish word *bajar*. In order to filter out the false pair (*subside, apear*) while keeping (*subside, baixar*), we compute the distributional similarity of those pairs using comparable corpora. As *subside* appears in very different contexts than the Galician word *apear* (*take down*), this bilingual pair is removed from the derived dictionary. Let us now suppose that we are building an English-Spanish list of bilingual terms by selecting those with high string similarity. This procedure is known as a bilingual cognates search. In the process of searching for bilingual cognates, the main problem that arises concerns false friends. For instance, the spelling of the English noun *code* is very close to that of the Spanish word *codo* (*elbow*). They are separated by an Edit Distance of only 1 (i.e. they differ by just one character). However, they cannot be considered to be translation candidates because their meanings are very different. To filter out fake bilingual pairs and select the correct ones, we again use their distribution similarity in comparable corpora. As both words do not appear in similar contexts, the pair is removed from the list of bilingual cognates.

In short, the specific objective of this chapter is to describe two methods used to derive new bilingual lexicons using comparable corpora to select correct candidate pairs. The first method consists of using two existing bilingual dictionaries, (*A, B*) and (*B, C*), in order obtain a new pair (*A, C*) by simple transitivity and, then, in validating correctly generated bilingual correspondences by using dependency-based distributional similarity computed from comparable corpora. The second method consists of generating candidate cognates from comparable corpora and, then, in validating correct candidates by computing their dependency-based distributional similarity in those corpora. As the experiments conducted will show, the performance of these strategies in terms of precision is close to the precision achieved by extraction methods based on parallel corpora.

This chapter organizes, integrates and expands the work presented in two previous articles: (Gamallo & Pichel 2010; Gamallo & Garcia 2012) with further experiments. The rest of the chapter is organized as follows: Section 2 describes the different steps underlying the method of building lexicons by transitivity with a pivot language, then, Section 3 describes the cognate-based strategy. In Section 4, we describe some experiments aimed at generating new bilingual dictionaries and evaluating the performance of our different strategies. Finally, some conclusions are presented in Section 5.

2. Pruning lexicons built through transitiviy

Our strategy[1] consists of two main tasks: both the generation of candidate bilingual correspondences by transitivity and their validation by using translation equivalents extracted from comparable corpora. This strategy is especially well-suited to creating new language resources for minority languages (e.g., Galician) from languages such as English or Spanish, which have many more resources. The method does not require the minority language to be provided with many large linguistic resources: only some raw text is required. This is enough to automatically build a new non-noisy, bilingual lexicon.

[1] The strategy was implemented in a prototype available at: <http://gramatica.usc.es/~gamallo/prototypes/BilingualExtraction.tar.gz>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

2.1 Basic assumptions

The crucial aspect of the method is the process of validating bilingual correspondences derived by transitivity, by means of translation equivalents extracted from comparable corpora. We observed that if a bilingual pair derived by transitivity also appears in the list of pairs extracted by distributional similarity from comparable corpora, then the pair is correct. This observation is supported by the following linguistic conjectures:

- In manually compiled bilingual dictionaries, each bilingual correspondence consists of two terms that share two different aspects of their lexical meaning: both of them have similar conceptual and distributional properties. It follows that the two terms both refer to similar entities or concepts (conceptual properties) and combine with similar entities or concepts (distributional properties).
- In noisy bilingual dictionaries derived by transitivity, most generated correspondences consist of bilingual pairs that share similar conceptual properties, but do not always have the same distributional properties. This is because the different senses of a polysemous word are related by conceptual aspects but not in distributional terms. Only homonymous words in the pivot language give rise to completely false and unrelated bilingual correspondences generated by transitivity.
- In bilingual lexicons automatically extracted from comparable corpora, the extracted correspondences consist of bilingual pairs with the same distributional properties, but which do not always share similar conceptual properties.

It follows that correct bilingual pairs are those that only share both conceptual and distributional properties. Then, the intersection of the dictionaries derived by transitivity (conceptual similarity) with those which are extracted from comparable corpora (distributional similarity) give rise to correct bilingual pairs, that is, to pairs that share both conceptual and distributional similarity. This intersection results in a bilingual, non-noisy lexicon.

The distributional hypothesis states that two words are semantically related if they share similar linguistic contexts. In a bilingual framework, this hypothesis may make it possible to identify translation candidates. The procedure works as follows: a word *w2* in the target language is a candidate translation of *w1* in the source language if the context expressions with which *w2* co-occurs tend to be translations of the context expressions with which *w1* co-occurs. The basis of the method is to find the target words that have the most similar distributions with a given source word. The starting point of this strategy is a list of bilingual expressions that are used to build the context vectors defining all words in both languages. This list is usually derived from an external bilingual dictionary.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

2.2 The pruning method

Let us look at the toy example in Figure 1. The objective is to generate non-noisy English–Portuguese pairs using Spanish as a pivot language.

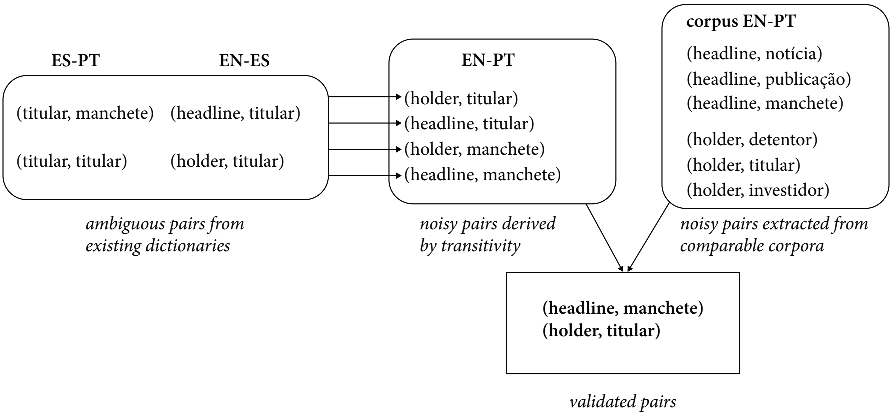


Figure 1. Example of the validation process

In the English–Spanish dictionary, the Spanish noun *titular* has two different bilingual correspondences: (*headline, titular*) and (*holder, titular*). This noun is then a polysemous word which also appears with two translations in the Portuguese–Spanish dictionary: (*titular, manchete*) and (*titular, titular*). The two senses of the Spanish polysemous word are conceptually related: the two refer to small text labels, the headline of an article or a person's name, used to identify either the specific article of a journal or a specific card owner.

The English–Portuguese pairs derived by transitivity are: (*headline, manchete*), *(*holder, manchete*), *(*headline, titular*), (*holder, titular*), where "*" stands for incorrect pairs. So, derivation by transitivity overgenerates bilingual pairs when one of the source words has multiple senses. According to our conjectures, even if all these generated pairs are somehow conceptually related, only those that are also distributionally similar can be considered to be correct. To identify the correct pairs, we make use of the translation equivalents extracted from an English–Portuguese comparable corpus, which allow us to validate those distributionally related pairs. In our experiments (described later in Section 4), the Portuguese translation candidates of *headline* extracted by our system are the following nouns: *notícia* (*news*), *publicação* (*publication*), *manchete* (*headline*), etc. These words are distributionally similar, but only *manchete* (in bold in Figure 1) describes the same concept as *headline*.

The noun *titular* (*holder*) was not extracted because its word distribution is very different to that of *headline*. On the other hand, the Portuguese translation

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

candidates of *holder* extracted by our systems are nouns like *detentor* (*detainer*), *titular* (*holder*), *investidor* (*investor*), etc. All of them have a similar distribution (agents of verb actions), but only the second one in Figure 1 (in bold) refers to the same concept as *holder*. The term *manchete* (*headline*) was not extracted because its distribution is very different from that of *holder*. This intersection results in a bilingual, non-noisy lexicon. So, the final intersection between the noisy pairs generated by transitivity and those derived from comparable corpora yields correct bilingual pairs.

It is worth mentioning that computing distributional similarity from comparable corpora requires some external bilingual resources to generate seed contexts. These seed contexts are conceived as anchors to (pseudo)-align the bilingual text corpora.

3. Pruning bilingual cognates

3.1 Basic assumptions

An efficient strategy used to build high quality bilingual lexicons between closely related languages is to search for bilingual cognates in highly comparable corpora. *Bilingual cognates* are considered here to be those words in two languages with similar spelling and similar meaning. There are, at least, three different aspects involved in the correctness of these kinds of lexicons:

Corpus similarity: The more comparable the corpora are, the more efficient the extraction performed on them is. In this way, we will discover similar articles in Wikipedia with very high degree of comparability (pseudo-parallel texts).

Distributional similarity: Words with similar distribution in comparable corpora are likely to be translation equivalents. As in the previous strategy (transitivity), we will use distributional similarity for validation, namely to validate the correct cognates.

Spelling similarity: Two words with the same or almost the same spelling are good candidates to be bilingual cognates. We will use the Edit distance to identify similar words in terms of spelling. To minimize the low coverage of the lexicons acquired using this method, it is convenient to conduct the experiments on a family of related languages. Indeed, only languages belonging to the same linguistic family share many cognates.

Only this third aspect (spelling similarity) is exclusive to cognate extraction. Distributional similarity is the validation method we used in the two proposed strategies (transitivity and cognates). Corpus comparability is an aspect that affects the two strategies, however, we only measure comparability in cognate extraction

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

because most cognates are technical terms appearing in domain specific texts. So, finding and selecting bilingual texts in the same technical domain is a crucial issue. The degree of comparability is very important in cognate extraction, but this is not enough to obtain non-noisy bilingual lexicons of cognates. The combination of the aforementioned three types of similarity, including distributional similarity, is required to generate high-quality cognate lexicons.

3.2 The pruning method

Our cognate-based method follows the idea that the use of distributional similarity to extract bilingual cognates from very comparable corpora should generate correct translations. Considering this idea, we designed a strategy adapted to the Wikipedia structure. Among the different web sources of comparable corpora, Wikipedia is likely to be the largest repository of similar texts in many languages. We only require the appropriate computational tools to make them comparable. The proposed method is based on the Wikipedia structure, even though it can be easily generalized to be adapted to other sources of comparable corpora.

The output is bilingual terminology containing many domain-specific terms found on Wikipedia. The method consists of four steps:

Corpus alignment: First, we identify the Wikipedia articles in two languages whose titles are translations of each other.

Degree of comparability: Then, to calculate a degree of comparability between two aligned articles, we apply a similarity measure and select the most comparable pairs of bilingual articles (Gamallo & González 2011b).

Candidates for translation equivalents: From each very comparable pair of articles, we calculate the distributional similarity and select the most similar word pairs, which are considered to be candidates for translation equivalents (Gamallo & Pichel 2008; Gamallo 2007). We also take multiword into account. Two bilingual dictionaries are required to generate lexico-syntactic seeds, which are used as text anchors in relation to both languages.

Selecting cognates: Finally, using Edit Distance (Levenshtein version), we check whether the candidates are *cognates* and select the most similar ones as true translation equivalents.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. Experiments

To verify whether our methods are useful, we used them to generate bilingual dictionaries in two different tasks. First, as was described in Section 2, two new bilingual dictionaries were built by transitivity: (*English, Galician*) and (*English, Portuguese*) dictionaries. Second, we produced bilingual terminology (*Spanish, Portuguese*) with bilingual cognates by making use of the strategy defined in Section 3. These resources were evaluated. All dictionaries and terminologies we produced are freely available.[2]

4.1 Derivation by transitivity

In this task, we built two new free dictionaries: (*English, Galician*) and (*English, Portuguese*). The existing dictionaries used as sources for deriving the new ones by transitivity are the following (only nouns, adjectives, and verbs are considered):

English–Spanish: For this pair, we used two resources: the free dictionary from Apertium v0.8,[3] which contains 10,828 bilingual entries, and the Collins[4] dictionary, which contains 48,637 entries.

Spanish–Portuguese: In this case, we used only the free resource Apertium v1.1,[5] which contains 10,281 entries.

Spanish–Galician: We also used Apertium v1.0,[6] which contains, for this pair of languages, 27,640 entries.

The Apertium dictionaries contain few multi-words, just some idioms. For this reason, we did not carry out multi-word extraction within this particular task. Using the strategy described above in Section 2, we generated the noisy bilingual dictionaries showed in Table 1 by transitivity.

Table 1. Noisy dictionaries derived by transitivity

Dictionaries	Total number	Ambiguous entries	Unambiguous entries
(English, Galician)	25, 790	18, 623	7, 167
(English, Portuguese)	12, 306	7, 179	5, 127

[2] <http://fegalaz.usc.es/~gamallo/dicos_comparable.tgz>

[3] <http://sourceforge.net/projects/apertium/files/apertium-en-es>

[4] <http://www.collinslanguage.com/>

[5] <http://sourceforge.net/projects/apertium/files/apertium-es-pt>

[6] <http://sourceforge.net/projects/apertium/files/apertium-es-gl>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Note that the third column of the table shows the number of ambiguous entries, which are actually the noisy entries. The words making up each entry pair are conceptually related even if many of them are not correct bilingual correspondences. The next step is to validate the noisy part of the dictionary by making use of translation equivalents extracted from comparable corpora, that is by making use of distributional similarity.

4.2 Comparable corpora

To validate the English–Galician and English–Portuguese correspondences with ambiguous words, we used the distributional-based strategy described in Gamallo (2007). Text corpora were syntactically analyzed using a multilingual dependency based parser, DepPattern (Gamallo & González 2011a). The comparable corpora were basically produced with news crawled from different online journals: 70*Mb* of English news from The New York Times,[7] Reuters Agency,[8] and The Guardian;[9] 70*Mb* of Galician news from Vieiros[10] and Galicia-Hoxe[11]; 21*Mb* of Portuguese news from Jornal de Notícias[12] and Público.[13] These monolingual corpora were used to build both (*English, Galician*) and (*English, Portuguese*) comparable corpora. Automatic extraction of translation equivalents were carried out on those comparable corpora by making use of the unambiguous entries generated by transitivity. These entries were used to generate lexico-syntactic seeds. Two sets of translation candidates were built: 700,000 candidates from (*English, Galician*) and 500,000 candidates from (*English, Portuguese*). Corpus-based lexicons are much bigger than those directly derived by transitivity because each word is associated to its *N* most appropriate translation candidates (where $N = 10$ in our experiments) by using distributional similarity. So, they contain much more noisy correspondences than those generated by transitivity. Ideally, they should contain, at least, a good bilingual correspondence for each word. This good correspondence will be used to validate dubious pairs derived by transitivity.

[7] <http://www.nytimes.com/>

[8] <http://trec.nist.gov/data/reuters/reuters.html>

[9] <http://www.theguardian.com/>

[10] <http://www.vieiros.com/>

[11] <http://www.galiciahoxe.com/>

[12] <http://www.galiciahoxe.com/>

[13] <http://www.publico.pt/>

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

4.2.1 *Validation*

To check the validity of the dubious correspondences within the ambiguity-based lexicons (i.e. containing ambiguous entries), we make their intersection with the corpus-based lexicons. This way, we filter out odd bilingual correspondences so as to just select the correct ones, which share both conceptual and distributional similarity.

The second column in Table 2 shows the validation number resulting from unifying both the distributional dictionaries and the ambiguous entries obtained by inter-section. In the (English, Galician) dictionary, we validated 4,248 correct entries which represent almost 23% of entries found in the ambiguity-based dictionaries (18,623 entries). In the (English, Portuguese) dictionary we validated 2,411 out of 7,179 ambiguous entries, which is 33.5% of all ambiguous entries generated by transitivity. These results are very similar to those obtained by Nerima and Wehrli (2008) using parallel corpora. These authors reported an experiment to derive an English–German dictionary by transitivity, where the ambiguity-based correspondences were validated using parallel corpora. The result of this checking process allowed them to validate 6,282 correspondences, which represent 26% of all candidate correspondences with ambiguous words. Even though we use non-parallel or comparable corpora, our results cannot be considered as being worse, which is very promising.

Table 2. Final non-noisy dictionaries

Dictionaries	Ambiguous (validated)	Unambiguous entries	Total entries
(English, Galician)	4, 248	7, 167	11, 415
(English, Portuguese)	2, 411	5, 127	7, 538

The third column in Table 2 shows the number of unambiguous entries, that is, those with one-to-one bilingual correspondences. Note that unambiguous entries are not required to be validated because they must all be correct considering that the manually compiled bilingual dictionaries taken as lexical source are also correct.

At the end of the process, the resulting non-noisy dictionary is the union of the validated correspondences with the lexicons containing unambiguous words. The last column shows the total number of non-noisy correspondences that our method was able to automatically generate. To be precise, we generated 11,415 entries in the (*English, Galician*) dictionary, which represent 44% of the total correspondences found in the original and noisy dictionary generated by transitivity (25,790). On the other hand, we generated 7,538 entries in the (*English, Portuguese*) dictionary, which represent 62% of the total correspondences found in the original and noisy dictionary generated by transitivity (12,206).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

4.2.2 *Evaluation of the dictionaries generated by transitivity*

To evaluate the correctness of the lexicons, we have selected several samples of 200 word pairs for each dictionary and for each subset of entries to be evaluated. More precisely, we evaluated the performance of both the list of unambiguous words as well as the process of validating ambiguous words with comparable corpora.

As expected, the set of unambiguous words generated by transitivity is 100% correct in terms of precision for both language pairs. No error was found.

As far as the validation process is concerned, Table 3 shows the results of our evaluation. Precision is the number of correct pairs validated by our system divided by all validated pairs. We found just two errors (99.0% precision) in the (*English, Galician*) sample and one error (99.5% precision) in the (*English, Portuguese*) one. Recall is the number of correct pairs validated by our system divided by all correct pairs found in the set of ambiguous entries. We estimated the number of correct ambiguous entries by using a sample of 200 ambiguous pairs for each language pair before validation. As we found that the percentage of correct ambiguous pairs is respectively 80% and 79% in the (*English, Galician*) and (*English, Portuguese*) dictionaries, the total number of correct ambiguous pairs likely to be extracted in (*English, Galician*) is 14,898, and 5,671 in (*English, Portuguese*). As Table 3 shows, the recall is still far from reasonable, in particular when the source dictionaries contain many ambiguous pairs that are not very frequent words, as in the case of (*Spanish, Galician*). This is in accordance with the results described in related work (Saralegi et al. 2011; Saralegi et al. 2012), where the authors provide good precision but recall is negatively affected. However, the correctness of the derived lexicons is similar to the dictionaries built by hand by lexicographers, since they are close to 100% correct. Moreover, in spite of the low recall, the size of the (*English, Galician*) dictionary is larger that the smaller source dictionary: namely, the *English –Spanish* lexicon integrated in the machine translation system Apertium (Armentano-Oller et al. 2006). It follows that our automatically generated dictionaries are both good and large enough to be inserted in rule-based machine translation systems.

Table 3. Evaluation of the validated bilingual correspondences

Dictionaries	Precision	Recall	F-score
(English, Galician)	99.0%	28.2%	43.8
(English, Portuguese)	99.5%	42.3%	59.3%

In fact, one of the direct applications of the two new generated dictionaries is their integration into an open source machine translation system: Apertium. More precisely, the main objective of our experiments is to update the bilingual lexicons of Apertium in order to improve the results of the machine translation system.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

4.3 Bilingual cognates

We conducted another experiment aimed at learning a large set of new bilingual cognates from the Portuguese and Spanish versions of Wikipedia. To minimize the low coverage of the lexicons acquired by the cognate-based method, it is convenient to use it on families of related languages, which share many cognates. This is why our experiments were carried out using Portuguese and Spanish, two Latin languages that are closely related. The extraction is focused on nouns, adjectives, and verbs, as well as on multi-words.

4.3.1 *Existing resources*

Our method requires a list of seed lexico-syntactic patterns, whose constituent lemmas are taken from existing bilingual resources. We used two different existing dictionaries:

- Apertium: The general purpose bilingual dictionary (*Spanish, Portuguese*) available in Apertium, and already used in the previous experiment. It contains 9,854 bilingual entries with nouns, adjectives, and verbs.
- Wikipedia: We created a new (Spanish, Portuguese) dictionary using the interlanguage links of Wikipedia. Since Wikipedia is an encyclopedia dealing with named entities and terms, this new dictionary only contains names and domain-specific terminology. It has up to 253,367 bilingual entries.

4.3.2 *Size of the extracted lexicons*

The total size reached by the union of both resources is 263,362 different bilingual correspondences, which will be used as seed pairs. Note that the two dictionaries are complementary: we only found 263 entries in common.

After applying our method to the whole Portuguese and Spanish Wikipedia, we extracted 27,843 new bilingual correspondences. None of them were in the two input dictionaries.

Table 4 depicts the final results. In the first row, we show the extractions of single words while the second row is focused on multi-words. Single words and multiwords are distributed by PoS categories: nouns, adjectives, and verbs. As far as multi-words are concerned, adjectives are not considered. The total extractions considering both multi-word terms and single words are shown in the third row. Notice that the total size of the new bilingual dictionary at 27,843 entries, is much larger than that of the general purpose dictionary of Apertium, which contains only 9,854 bilingual correspondences.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4. Size of the extracted lexicons

	Nouns	Adjectives	Verbs	Total
Single words	9,374	5,725	2,215	17,314
Multi-word terms	9,585	-	944	10,529
All terms	18,959	5,725	3,159	27,843

4.3.3 *Evaluation of the cognate-based extraction*

To evaluate the precision of the extracted dictionary, a test set of 450 bilingual pairs was randomly selected, consisting of three balanced subsets: 150 bilingual pairs of nouns, 150 bilingual pairs of verbs, and 150 bilingual pairs of adjectives. These included nominal and verbal multi-words, where their head is either a noun or a verb. The results are depicted in Table 5. Precision is the number of correct pairs divided by the number of evaluated pairs. Recall is the number of correct pairs divided by the number of all correct candidates extracted before the final validation performed with distributional similarity. So, we consider that the total number of correct candidates is that provided by distributional similarity, just before being validated with the cognate-based strategy. To compute recall, we took into account the number of correct pairs extracted by the distributional similarity that were not selected by the cognate-based similarity. For this purpose, we used a new test set with 200 pairs (separated by PoS categories). We found that only 9% of those pairs which were not validated were correct. It follows that our cognate-based similarity is losing few correct cases, giving rise to high recall.

Table 5. Evaluation of the extracted bilingual cognates

	Precision	Recall	F-score
Nouns	91.0%	88.7%	89.5
Verbs	89%	86.8%	87.9%
Adjectives	95.9%	94.8%	94.5.9%
Total	92%	89.5%	90.7%

The best performance was achieved by using adjectives: 95% precision and 94% f-score. By contrast, verb extraction only achieves 89% precision. The performance for adjectives is better than that for verbs and nouns probably because adjectives are not on the list of multi-words.

The precision of the total bilingual lexicon, with 27,843 entries, is 92%. This performance outperforms state-of-the-art work on extraction from comparable corpora, whose best scores were about 70% accuracy in Rapp (1999) and 60-83% in Aker et al. (2013). The correctness of the generated translation equivalents is

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

similar to that achieved using parallel corpora. It follows that our method permits the minimization of the effort to build a new bilingual dictionary of two related languages.

It is worth mentioning that the results of the comparability measure have not been directly evaluated. In order to know the error rate underlying the automatic alignment of similar texts, a quantitative analysis will be required in future work.

4.3.4 *Error analysis*

We found 39 errors out of 450 evaluated extractions. Most of them (58%) were due to foreign words, namely English words appearing in the input text as part of titles or citations. For instance, the translation pair "about/about" was incorrectly learned from two Portuguese and Spanish texts containing such a word within a non-translated English expression. It would not be difficult to avoid this kind of problem if we use automatic language identification to find parts of the input text written in other languages.

The second most common error type (8%) was caused by prefixes appearing in one of the two correlated words, for instance:

1. americanismo/anti-americanismo (anti-americanism)
2. anti-fascista/fascista (anti-fascist)
3. hispanoárabe/neo-hispano-árabe (neo-hispano-arabic)

Note that it would be possible to filter out those cases by making use of a list of productive prefixes.

In Table 6, we show some types of errors found in the evaluation. As the two most common errors (foreign words and prefixes), which represent 66% of the total number of errors, can be easily filtered out, the total achievable accuracy of our system could be 97%.

Table 6. Types of errors ranked by frequency

Error types	Frequency (%)
foreign words	58%
prefixes	8%
typos	8%
multi-words	5%
POS tagging	3%

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

5. Conclusions and future work

In this chapter, we described two different methods to build high quality bilingual lexicons using comparable corpora. In order to overcome the poor results and low precision inherent to most extraction approaches based on comparable corpora, we made use of two restrictions: transitivity and cognates. The performance of our approach, in terms of precision, is close to the precision achieved by the extraction methods based on parallel corpora.

We made use of dictionaries already integrated into rule-based machine translation systems such as Apertium. It follows that an application of our method will be helpful for the production of new language pairs treated by a machine translation system, namely those pairs included in minority languages. Further evaluations of the results obtained with machine translation systems could be considered as an indirect evaluation of the correctness of the dictionaries produced by our extraction strategies. The number of bilingual dictionaries required by a multilingual translator increases as a quadratic function of the number of languages the system aims to translate (Wehrli et al. 2009). Hence, the process of automatically deriving new bilingual resources can drastically reduce the amount of work required for this task. Moreover, as our extraction methods only require comparable corpora, it will not be difficult to generate new bilingual dictionaries and terminologies for those languages with fewer resources or with fewer parallel texts available.

Acknowledgements

This work has received financial support from Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

Aker, Ahmet, Paramita, Monica & Gaizauskas, Robert. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 4–9, Sofia, Bulgaria.

Ansari, Ebrahim, Sadreddini, Mohammad H., Tabebordbar, Alireza & Mehdi, Sheikhalishahi. 2014. Combining different seed dictionaries to extract lexicon from comparable corpus. *Indian Journal of Science and Technology*, 7(9):1279–1288.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Armentano-Oller, Carme, Carrasco, Rafael C., Corbí-Bellot, Antonio M., Forcada, Mikel L. Mireia, Rosell, Ginestí, Ortiz-Rojas, Sergio, Pérez-Ortiz, Juan Antonio, Ramírez-Sánchez, Gema, Sánchez-Martínez, Felipe & Scalco, Miriam A. 2006. Open-source Portuguese–Spanish machine translation. *Lecture Notes in Computer Science* 3960, 50–59. https://doi.org/10.1007/11751984_6

Gamallo, Pablo. 2007. Learning bilingual lexicons from comparable English and Spanish corpora. In *Machine Translation SUMMIT XI*, 191–198, Copenhagen, Denmark. <https://pdfs.semanticscholar.org/5776/8274c94e730a92c5606bd7d7703da12146da.pdf>

Gamallo, Pablo & Garcia, Marcos. 2012. Extraction of bilingual cognates from Wikipedia. *Lecture Notes in Computer Science* 7243: 63–72. https://doi.org/10.1007/978-3-642-28885-2_7

Gamallo, Pablo & González, Isaac. 2011a. A grammatical formalism based on patterns of part-of-speech tags. *International Journal of Corpus Linguistics* 16(1): 45–71. https://doi.org/10.1075/ijcl.16.1.03gam

Gamallo, Pablo & González, Isaac. 2011b. Measuring comparability of multilingual corpora extracted from Wikipedia. In *Workshop on Iberian Cross-Language NLP tasks (ICL-2011)*, 8–13, Huelva, Spain.

Gamallo, Pablo & Pichel, José Ramom. 2008. Learning Spanish–Galician translation equivalents using a comparable corpus and a bilingual dictionary. *Lecture Notes in Computer Science* 4919: 413–423.

Gamallo, Pablo & Pichel, José Ramón. 2010. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. In *CICLING, LNCS*, Vol. 6008, 473–483, Iasi, Romania. Heidelberg: Springer.

Hazem, Amir & Morin, Emmanuel. 2014. Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models. *Lecture Notes in Computer Science* 8404: 310–323. https://doi.org/10.1007/978-3-642-54903-8_26

Nakagawa, Hiroshi. 2001. Disambiguation of single noun translations extracted from bilingual comparable corpora. *Terminology* 7(1), 63–83. https://doi.org/10.1075/term.7.1.06nak

Nerima, Luka & Wehrli, Eric. 2008. Generating bilingual dictionaries by transitivity. In *LREC'08*, 2584–2587, Marrakesh, Marocco.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German Corpora. In *Proceedings of ACL'99*, 519–526.

Saralegi, Xabier, Manterola, Iker & San Vicente, Iñaki. 2011. Analyzing methods for improving precision of pivot-based bilingual dictionaries. In *Empirical Methods in Natural Language Processing (EMNLP-2011)*, 846–856, Edinburgh, Scotland, UK.

Saralegi, Xabier, Manterola, Iker & San Vicente, Iñaki. 2012. Building a Basque-Chinese dictionary by using English as pivot. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, 1443–1447, Istanbul, Turkey.

Tamura, Akihiro, Watanabe, Taro & Sumita, Eiichiro. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 24–36, Jeju Island, Korea.

Wehrli, Eric, Nerima, Luka & Scherrer, Yves. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *4th Workshop on Statistical Machine Translation*, 90–94, Athens, Greece.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Discovering bilingual collocations in parallel corpora

A first attempt at using distributional semantics

Marcos Garcia, Marcos García-Salido and
Margarita Alonso-Ramos
University of A Coruña

This chapter presents a method that exploits parallel corpora to automatically extract bilingual collocation equivalents. First, we use dependency parsing and statistical measures to identify collocation candidates in corpora. Then, we leverage the parallel corpora to extract bilingual word-embeddings. Finally, we use these distributional models as probabilistic dictionaries in order to identify bilingual collocation equivalents. To evaluate our strategy we carry out a set of experiments in Portuguese and Spanish focusing on verb-object collocations, for example, "reach the maturity" ("atingir a maturidade" in Portuguese, "alcanzar la madurez" in Spanish). The results of our experiments show that this method is useful to automatically identify thousands of bilingual collocation equivalents, achieving a precision of 86%.

Keywords: collocations, distributional semantics, phraseology, parallel corpora

1. Introduction

When learning a new language, non-native speakers usually face difficulties regarding the use of conventionalized lexical combinations in each linguistic variety (Altenberg & Granger 2001). It is common for language learners to produce false combinations, usually influenced by their mother tongue (Nesselhauf 2004; Alonso-Ramos et al. 2010). One such type of expression is *collocations*, which can be defined as unpredictable lexically restricted combinations of two lexical units (Mel'čuk 1998). As an example, a native speaker of Portuguese might use **give a walk* instead of *take a walk* in English, guided by the Portuguese collocation *dar um passeio,* since the primary meaning of the verb *dar* corresponds to *to give*.

https://doi.org/10.1075/scl.90.16gon

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

© 2019 John Benjamins Publishing Company

The manual creation of large structured lists of bilingual collocation equivalents involves a great deal of effort from expert lexicographers. Nevertheless, these resources could be very useful for a variety of activities such as second language learning or tasks such as machine translation (Orliac & Dillinger 2003). In this respect, parallel corpora are interesting resources for the extraction of crosslinguistic equivalents and have been widely used for identifying both monolexical equivalents and plurilexical entries (named entities, multiword expressions, etc.) (Fung 1998; Smadja 1993; Şulea et al. 2016).

In this chapter, we explore the use of parallel corpora not only to discover bilingual collocation equivalents, but also to create a bilingual distributional semantics model. This model is used to extract, with a high degree of precision, bilingual collocation equivalents from corpora.

In the proposed approach, we follow the phraseological criterion that defines a collocation as a restricted pair of lexical units (LUs), where one of them (the *base)* is freely selected by the speaker (e.g., *walk* in *take a walk*), while the other one (the *collocate*; for example, *take* in *take a walk*) is restricted by the base (Mel'čuk 1998). Thus, our strategy for discovering bilingual collocation synonyms consists of searching for semantically similar equivalents of both the base (in a first step) and the collocate of each collocation. As the model only makes use of distributional information, it is worth noting that, ideally, it could be applied to other non-parallel corpora for extracting additional examples of collocation equivalents not present in the original bilingual resource.

In order to test the proposed strategy, we carried out an experiment to automatically discover bilingual *verb-object* collocations in a Spanish–Portuguese corpus. The results of this test show that this method enables high-quality bilingual collocation equivalents to be extracted with no manual effort whatsoever.

2. Previous research on bilingual collocation extraction

Since the 1990s, a number of works have exploited parallel corpora to extract bilingual collocation equivalents. Smadja (1992) and Smadja et al. (1996) use a parallel English–French corpus first to identify monolingual collocations in a source language (English), and then to discover the French equivalents of the source collocations by applying different similarity measures such as *Mutual Information* or *Dice Coefficient*.

Kupiec (1993) also uses the same parallel corpus to find English–French noun phrase equivalents. This method applies an expectation maximization algorithm on monolingual collocations that had previously been extracted.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Wu and Chang (2003) also take advantage of parallel corpora to extract Chinese and English *n-grams* from aligned sentences, by computing their *log-likelihood* ratio. Then, they apply a competitive linking algorithm to decide if the bilingual pairs are real Chinese–English translations.

Using syntactic analysis, Seretan and Wehrli (2007) extract bilingual collocations from parallel corpora. First, the authors obtain monolingual collocations using *log-likelihood*, and then they search for equivalents of each base using bilingual dictionaries. Although this is not always the case, this method assumes that bilingual collocation equivalents occur in the same syntactic relation in both the source and the target languages.

More recently, Rivera et al. (2013) present different methods for bilingual collocation extraction that can be applied to both parallel and comparable corpora. This work relies on *n-grams* to extract monolingual collocations, and on bilingual dictionaries (or WordNet) to identify the bilingual equivalents.

A different approach was adopted by Lü and Zhou (2004), who use non-related monolingual corpora for finding bilingual collocations. The work they present first applies dependency parsing and the *log-likelihood* ratio for discovering English and Chinese monolingual collocations. They subsequently search for bilingual collocations using translation equivalents – with the expectation maximization algorithm and bilingual dictionaries – of words that occur in the same dependency relation in both languages.

Most of these papers use contextual information (e.g., the position of the collocations in the corpus) to find the bilingual equivalents, together with bilingual dictionaries. In the present chapter, we introduce a different approach which eliminates the need for bilingual dictionaries and does not make use of explicit contextual information, making it easy to extend to other languages and applicable to different corpora. In Garcia et al. (2017) we presented the results of a similar research project using both parallel and comparable corpora in Portuguese, Spanish, and English.

3. The proposed strategy

This section presents a method for discovering bilingual collocations within corpora. The strategy consists of the following phases: First, we extract monolingual collocation candidates from corpora, using dependency parsing to identify syntactically related words together with the application of statistical measures for ranking the extracted candidates. Then, we train a bilingual distributional semantics model using a word level automatic alignment of the parallel corpora. Finally, this model is applied to the monolingual collocations to identify bilingual equivalents

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

automatically. Both the monolingual collocations and the distributional model are trained using lemmas instead of tokens, so that the system groups together different occurrences of the same collocation into its canonical form. For example, expressions such as *took several pictures*, *take one more picture*, *pictures were taken*, etc. are grouped together under the collocation "*picture*$_{Base}$, *take*$_{Collocate}$".

3.1 Extracting monolingual collocation candidates

As mentioned, we use dependency parsing to extract monolingual collocations candidates, enabling us to avoid the incorrect identification of non-related words that frequently co-occur in corpora (Evert 2008). Moreover, syntactic analysis also makes it possible to discover word combinations that appear in a relatively large span of text, which are not usually obtained using extraction methods based on *n-grams* (Smadja 1993).

Before extracting the collocation candidates, we enrich the corpus with linguistic information by means of automatic tokenization, lemmatization and POS tagging. Then, we apply a dependency parser to syntactically annotate the text using *Universal Dependencies* (Nivre et al. 2016). Universal Dependencies is a recent initiative aimed at providing treebanks in many languages labeled with a uniform annotation. The use of the same criteria for labelling the syntactic information facilitates a multilingual work such as the one presented in this chapter. Table 1 shows a representation of the sentence "She hates black money". In this representation, the first column represents the position of each token in the sentence, followed by its form, lemma and POS tags (columns 2 to 4). The last two columns indicate, respectively, the syntactic head of each token (where 0 means that this is the root of the sentence) and the dependency relation both tokens bear in this context (the dependent and the head).

After annotating the corpus, we extract candidates of different types of collocations: *verb-object* ("*statement*$_B$, *make*$_C$"), *adjective-noun* ("*money*$_B$, *black*$_C$"), and

Table 1. Example of a labeled sentence

id	token	lemma	POS tag	head	dep
1	She	she	PRON	2	nsubj
2	hates	hate	VERB	0	root
3	black	black	ADJ	4	amod
4	money	money	NOUN	2	dobj

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 2. Example of Portuguese *verb-object* candidates ranked by MI

ase	collocate	MI
pastilha	mascar	11.6
formulário	preencher	11.3
homenagem	render	10.9
susto	pregar	10.6
isco	morder	10.1

noun-(preposition)-noun ("*cigarette*$_B$, *packet*$_C$").[1] These candidates are then ranked using standard association measures such as *Mutual Information* (MI, which promotes low-frequency candidates and works well in large corpora (Pecina 2010)), *t-score* (which assigns high scores to frequent word combinations (Krenn & Evert 2001), or others such as *z-score* or *log-likelihood*. Table 2 shows examples of candidate collocations in Portuguese ranked by their mutual information values.

Thus, we obtain several lists (per language, per collocation type, and per association measure), which are then merged into a unique file per language and collocation type based on a given threshold.

3.2 Bilingual distributional semantics model

A (monolingual) model of distributional semantics maps each word of a vocabulary into a vector of real numbers which represents the distributional properties of the word. Recently, distributional semantics has become very popular in the natural language processing (NLP) community due to the publication of *word-2vec* (Mikolov et al. 2013), which uses neural network approaches to reduce the dimensionality of the vectors (*word-embeddings*). In a monolingual scenario, we can compute the cosine distance between the vectors of two words, obtaining a similarity measure between them.

In our case, the hypothesis behind the use of a bilingual model of distributional semantics is that it can effectively learn words with a similar distribution (in a target language) semantically related to those in the source language. Thus, as an independent task to the extraction of monolingual candidates, we train a bilingual model of distributional semantics using a naïve approach that takes advantage of an automatic alignment, at word level, of parallel corpora.

[1] Note that each collocation type has both syntactic and morphosyntactic restrictions: verb-object, for instance, would require a noun (as the base) and a verb (as the collocate) occurring in a dobj dependency relation.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

In order to rapidly build a bilingual model, we use the following strategy: As a prior step to training the bilingual model, we replace each token of the corpus by its lemma with a view to reducing data sparseness and making the corpus compatible with the monolingual collocations (which had been extracted using the lemmas). We also incorporate a language suffix, in order to better identify the language of each lemma (Figure 1).

Tirou várias fotografias	sacó varias fotografías	took several pictures
tirar$_{\text{PT}}$ vário$_{\text{PT}}$ fotografia$_{\text{PT}}$	sacar$_{\text{ES}}$ vario$_{\text{ES}}$ fotografía$_{\text{ES}}$	take$_{\text{EN}}$ several$_{\text{EN}}$ picture$_{\text{EN}}$

Figure 1. Example of original (top) and lemmatized and suffixed sentences (bottom) in Portuguese (left), Spanish (center), and English (right)

Then, we apply an automatic aligner to identify the correspondence of each word of the source language in the target one (Dyer et al. 2013). We use the information provided by the word alignment to create a *mixed bilingual* corpus by concatenating each equivalent lemma in the source and target languages:

(1) "tirar$_{\text{PT}}$ sacar$_{\text{ES}}$ vário$_{\text{PT}}$ vario$_{\text{ES}}$ fotografia$_{\text{PT}}$ fotografía$_{\text{ES}}$"

Thus, we obtain a mixed corpus with bilingual word equivalents occurring together (or close to each other), in order to train a naïve bilingual distributional model. To create the model, we apply *word2vec* in this final corpus using the *skip-gram* architecture, which assigns a higher weight to words nearby in the context than to more distant ones. We defined the vector dimension as 300 and used a context window of 20 tokens (≈10 in each language). Furthermore, we created distributional vectors only for those words with at least 5 occurrences in the corpus.

The resulting model contains, for each word in the corpus (in Portuguese and Spanish), its distributional vector which encodes information about the linguistic contexts in which the word occurs. Thus, given two words, the distance between their vectors (computed by their cosine distance) will ideally reflect their semantic similarity.

Using this bilingual model, we can search for the most similar words (in terms of their distribution) in a target language given a certain input from a source language. The plot in Figure 2 shows a simple 2D visualization example of a bilingual Spanish–Portuguese model.

Each point represents a word vector (whose size was automatically reduced to two dimensions using t-SNE (Maaten & Hinton 2008)), and the distance between two points corresponds to the semantic distance between the words they represent. In this example, we can see that the circled words such as *perfil* or *letra* (*profile* and *character* in Spanish and Portuguese, respectively) occupy (almost) the same vector space in both Spanish and Portuguese. However, other words have

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

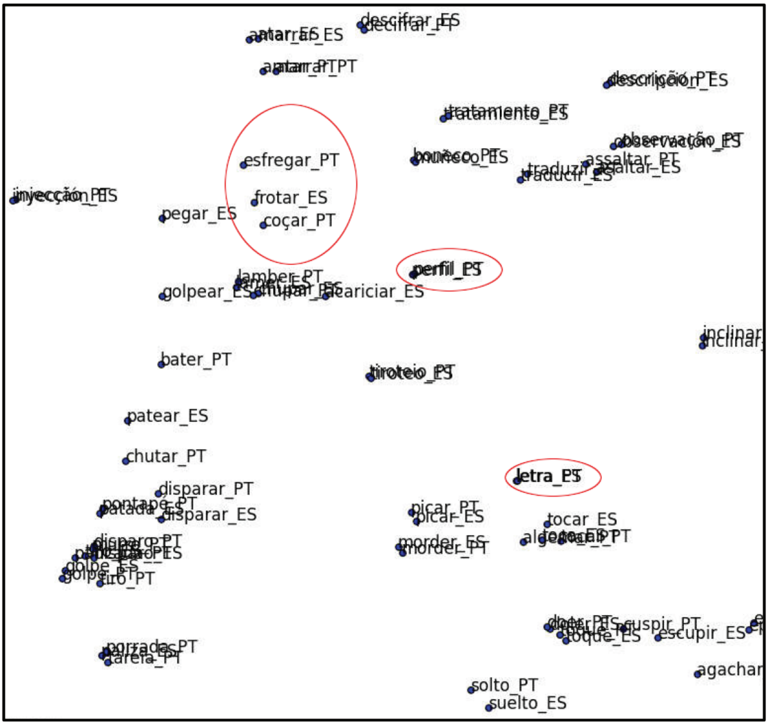


Figure 2. Visualization example of a two-dimensional Portuguese–Spanish model. Circles were manually included to facilitate the interpretation

different equivalents in the two languages: *frotar* (*to rub*, in Spanish) is very similar to both *coçar* (*to scratch*) and *esfregar* (*to scrub*) in Portuguese. Bilingual models of this kind will be used as probabilistic dictionaries to find equivalents of both the base and the collocate of the monolingual collocations.

3.3 Bilingual alignment of monolingual collocations

As pointed out, the distributional model is used to search for bilingual equivalents in the previously extracted monolingual collocations of the source language (*lang_A*) and in the target one (*lang_B*). We applied the following strategy:

- First, the collocations are traversed starting from those with a higher association score in *lang_A*.
- For each one (e.g., "*autobús*$_B$, *coger*$_C$" –*to take the bus* in Spanish), we select its base and search for equivalents in *lang_B* using the distributional model (e.g., *autocarro*, *camioneta*, *comboio*, etc., in Portuguese).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

- For words in *lang_B* with a similarity higher than a given threshold (*thres_base*), we verify whether collocations with these words as base are present in the list of *lang_B* (e.g., "*comboio*$_B$, *apanhar*$_C$", "*autocarro*$_B$, *apanhar*$_C$", "*comboio*$_B$, *perder*$_C$", etc.).
- If this is the case, we calculate the cosine distance between the collocates of the bilingual equivalents (*coger* versus *apanhar* and *coger* versus *perder*).
- Finally, we select a collocation equivalent if the similarity between the collocates is also higher than a predefined threshold (in this case, *thres_coll*). In the previous example, the Spanish collocation "*autobús*$_B$, *coger*$_C$" will be aligned with the Portuguese "*autocarro*$_B$, *apanhar*$_C$".

Note that "*comboio*$_B$, *apanhar*$_C$" (*to take the train*, in Portuguese) could also be aligned to "*autobús*$_B$, *coger*$_C$" using the proposed method. In order to decide the best candidate among those in one of such sets, we compute the average of the distances between both bases and collocates (in *lang_A* and *lang_B*) and use the resulting number as a confidence value for each pair of aligned collocations. Thus, "*autocarro*$_B$, *apanhar*$_C$" will have a confidence value of 0.85 regarding the source collocation, while "*comboio*$_B$, *apanhar*$_C$" will have a score of 0.78.

4. Evaluation

To test the method proposed in the previous section, we performed a preliminary evaluation extracting *verb-object* bilingual collocations from a Spanish (*es*)–Portuguese (*pt*) corpus.

4.1 Data

We used a part of the Spanish–Portuguese parallel corpus of OpenSubtitles2016 (Lison and Tiedemann 2016) for extracting the monolingual (*es* and *pt*) collocations, and to learn the bilingual word-embeddings. This corpus contains sentences from movies and TV series subtitles, the alignment of which is determined by the time they occur in the movie, so it is roughly similar to a sentence level alignment.

To extract the monolingual collocations we used about 2 million sentences, obtaining more than 150,000 verb-*dobj*-noun pairs in Spanish, and almost 215,000 in Portuguese.

To train the distributional model we used the proposed strategy on the first 11 million sentences of the corpus. The data were processed using LinguaKit for tokenizing, lemmatizing and POS tagging (Garcia & Gamallo 2015), and MaltParser to perform the dependency parsing (Nivre et al. 2007). The MaltParser models

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

were previously trained on the Portuguese and Spanish Universal Dependencies treebanks (version 1.3).

4.2 Monolingual extraction and bilingual alignment

The verb-*dobj*-noun extractions were ranked using mutual information. We then defined a threshold of *MI* = > *3* (and a frequency of *f* = > *6*), obtaining > 1,000 collocations with highest MI values (1,024 in Spanish and 1,059 in Portuguese).

From these two sets of ≈ 1,000 collocations we applied our method to identify bilingual equivalents, empirically defining *thres_base* and *thres_coll* as 0.65. In those cases where the system extracted more than one equivalent, we selected the most reliable one. The proposed strategy obtained 483 collocation equivalents out of ≈ 1,000 input collocations.

4.3 Results

We performed an initial evaluation of both the monolingual extraction and the bilingual alignment processes. To do so, we randomly selected 50 pairs of collocation equivalents and evaluated them regarding their (i) *collocability* (whether the combination could be actually classified as a phraseological collocation), and their (ii) bilingual equivalence (whether they could be used as translations in a real scenario). The evaluation was performed by two of the authors, while the third checked the dubious cases marked by the other reviewers.

The results of these *verb-object* collocation extractions achieved an average of 74% collocability, and 86% bilingual equivalence. In this regard, it is worth noting that some of the bilingual equivalents extracted by the system were marked as incorrect even though they appeared in real translations in the original corpus (e.g., "es:*dar un beso* = pt:*dar um abraço*" – *to kiss* in Spanish, and *to hug* in Portuguese). These cases may behave as translations in some contexts, but they were not labeled as semantically equivalent by the reviewers. If we had labeled these equivalents as correct, this process would have attained 92% precision.

4.4 Error analysis

Although the evaluation of our strategy is at an initial stage, we have carried out a brief error analysis aimed at discovering the main sources of errors in the bilingual alignment of monolingual collocations. The errors were classified in four different groups:

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

1. NLP tools: Containing errors produced by the NLP tools used for analyzing the corpus. For example, *loco* in *volver loco* (*to turn mad*, in Spanish) was incorrectly analyzed as a *dobj.*
2. Distributional semantics: A frequent issue in using distributional semantics approaches for finding synonyms (both monolexical and plurilexical) is that antonyms often occur in very similar contexts. In this regard, our strategy aligned (as bilingual equivalents) some collocations containing antonyms of the bases or of the collocates, for example, *"es:*esposas*$_B$*, *poner*$_C$ = pt:*algemas*$_B$*, *tirar*$_C$" (*to put the handcufffs on* in Spanish, and *to remove the handcuffs*, in Portuguese).
3. Association measures: The association measures extracted several combinations that are not phraseological collocations. For example, "*roupa*$_B$*, *comprar*$_C$" (*to buy clothes*, in Portuguese), even though the method identified bilingual equivalents correctly (e.g., "*ropa*$_B$*, *comprar*$_C$" in Spanish).
4. Corpora: Even if OpenSubtitles2016 is a useful resource for bilingual research, some translations are not particularly appropriate, yielding to the extraction of undesired equivalents. Also, several subtitles belonging to a specific variety appear in another variety of the same languages (for instance, many collocations extracted from the European Portuguese data were actually Brazilian combinations).

5. Conclusions

This chapter presents a method for automatically extracting bilingual collocation equivalents from parallel corpora. First, we extract monolingual collocation candidates using dependency parsing. Then, we train a naïve bilingual model of distributional semantics. Finally, this bilingual model is used to find equivalents of both the base and the collocate of the monolingual collocations.

Preliminary results of *verb-object* collocation extraction in Spanish and Portuguese show that this strategy can effectively discover bilingual collocation equivalents with 86% precision. Thus, using the proposed method it is feasible to automatically obtain large lists of bilingual collocation equivalents from corpora, which in turn can be useful resources for different tasks such as machine translation as well as the compilation of teaching material for foreign languages.

Our method does not require any information about the position of the collocation in the original corpus, this being the reason why it could be applied to different resources such as comparable corpora or even non-related texts. In this respect, further work will focus on testing this strategy in comparable corpora (such as those used in Gamallo 2019) and with different collocation patterns (such

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

as *noun-noun, adjective-noun* or *subject-verb* collocations), extending the work presented in Garcia et al. (2017). Moreover, it will be interesting to apply this method to less closely related languages.

Finally, the error analysis performed allowed us to detect some of the main sources of errors that should also be taken into account in further research.

Acknowledgements

This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through projects FFI2016-78299-P and FFI2014-51978-C2-1-R, by two *Juan de la Cierva* grants (FJCI-2014-22853 and IJCI-2016-29598), and by a postdoctoral fellowship endowed by the Galician Government (POS-A/2013/191)

References

Alonso-Ramos, Margarita, Wanner, Leo, Vincze, Orsolya, Casamayor del Bosque, Gerard, Vázquez Veiga, Nancy, Mosqueira Suárez, Estela & Prieto González, Sabela. 2010. Towards a motivated annotation schema of collocation errors in learner corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 3209–3214. Paris: European Language Resources Association (ELRA).

Altenberg, Bengt & Granger, Sylviane. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22: 173–195. https://doi.org/10.1093/applin/22.2.173

Dyer, Chris, Victor Chahuneau & Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2013), 644-648. Atlanta, Georgia: Association for Computational Linguistics.

Evert, Stefan. 2008. Corpora and collocations. In *Corpus Linguistics. An International Handbook*, Vol. 2, Anke Lüdeling & Merja Kytö (eds), 1212–1248. Berlin: Mouton de Gruyter.

Fung, Pascale. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas. Machine Translation and the Information Soup (AMTA 1998)*, 1–17, Langhorne, PA: Association for Machine Translation in the Americas.

Gamallo, Pablo. 2019. Strategies to build high quality bilingual lexicons from comparable corpora. In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* [Studies in Corpus Linguistics 90], Irene Doval & M. Teresa Sánchez (eds). Amsterdam: John Benjamins. (this volume)

Garcia, Marcos & Gamallo, Pablo. 2015. Yet another suite of multilingual NLP Tools. In *Languages, Applications and Technologies* [Communications in Computer and Information Science 563], José-Luis Sierra-Rodríguez, José Paulo Leal & Alberto Simões (eds), 65–75. Cham: Springer. Revised Selected Papers of the Symposium on Languages, Applications and Technologies (SLATE 2015), Madrid. https://doi.org/10.1007/978-3-319-27653-3_7

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Garcia, Marcos, Marcos García-Salido & Margarita Alonso-Ramos. 2017. Using bilingual word-embeddings for multilingual collocation extraction. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) at the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, 21-30. Valencia: Association for Computational Linguistics.

Krenn, Brigitte & Evert, Stefan. 2001. Can we do better than frequency? A case Study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, 39–46. Toulouse: Association for Computational Linguistics.

Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL 1993)*, 17–22, Columbus OH: Association for Computational Linguistics. https://doi.org/10.3115/981574.981577

Lison, Pierre & Tiedemann, Jörg. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 923–929. Paris: European Language Resources Association (ELRA).

Meurer, Paul. 2012. INESS-Search: A search system for LFG (and other) treebanks. In Proceedings of LFG12 Conference, Miriam, Butt & King, ButtTracy H. (eds). Stanford, CA: CSLI Publications).

Lü, Yajuan & Zhou, Ming. 2004. Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, 167–174. Barcelona: Association for Computational Linguistics.

van der Maaten, Laurens & Hinton, Geoffrey. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605.

Mel'čuk, Igor. 1998. Collocations and lexical functions. In *Phraseology. Theory, Analysis and Applications*, Anthony Paul Cowie (ed.), 23–53. Oxford: Clarendon Press.

Mikolov, Tomas, Chen, Kai, Corrado, Greg & Dean, Jeffrey. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR 2013)*. Scottsdale AZ. ArXiv preprint arXiv:1301.3781.

Nesselhauf, Nadja. 2004. *Collocations in a Learner Corpus* [Studies in Corpus Linguistics 14]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.14

Nivre, Joakim, de Marneffe, Marie-Catherine, Ginter, Filip, Goldberg, Yoav, Hajič, D. Manning, Christopher, McDonald, Ryan, Petrov, Slav, Pyysalo, Sampo, Silveira, Natalia, Tsarfaty, Reut & Zeman, Daniel. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1659–1666. Paris: European Language Resources Association (ELRA).

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandra, Marinov, Svetoslav & Marsi. Erwin. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(2): 95–135.

Orliac, Brigitte & Dillinger, Mike. 2003. Collocation extraction for machine translation. In *Proceedings of Ninth Machine Translation Summit (MT Summit IX)*, 292–298, New Orleans LA.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1-2): 137–158. https://doi.org/10.1007/s10579-009-9101-4

Rivera, Oscar Mendoza, Mitkov, Ruslan & Corpas Pastor, Gloria. 2013. A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In *Multiword Units in Machine Translation and Translation Technology* [Current Issues in Linguistic

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Theory 341], Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor & Violeta Seretan (eds), 18–25, Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.341.08riv

Seretan, Violeta & Wehrli. Eric. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conference sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, 401–410, Toulouse.

Smadja, Frank, McKeown, Kathleen R. & Hatzivassiloglou, Vasileios. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22(1): 1–38.

Smadja, Frank. 1992. How to compile a bilingual collocational lexicon automatically. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, 57–63, San Jose CA.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational linguistics* 19(1): 143–177.

Şulea, Octavia-Maria, Nisioi, Sergiu & Dinu, Liviu P. 2016. Using word embeddings to translate named entities, In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3362–3366. Paris: European Language Resources Association (ELRA).

Wu, Chien-Cheng & Chang, Jason S. 2003. Bilingual collocation extraction based on syntactic and statistical analyses. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing (ROCLING 2003)*, 1–20. Taiwan: Association for Computational Linguistics and Chinese Language Processing.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Normalization of shorthand forms in French text messages using word embedding and machine translation

Parijat Ghoshal and Xi Rao
Neue Zürcher Zeitung, KOF Swiss Economic Institute

This chapter focuses on the normalization of abbreviations and shorthand forms used in French text messages. These forms are difficult to normalize, as they mostly cannot be resolved by typical spell checkers and dictionary lookups. Firstly, we aligned normalized and non-normalized French text messages and built a parallel corpus. We applied two popular approaches for text normalization, namely multilingual word embeddings, and character-based machine translation. We compare our results and observe the efficacy of our models while normalizing deletions, substitutions, repetitions, swaps, and insertions, made to canonical forms. This is the first paper that uses Multivec and the Belgian SMS corpus collected under the SMS4Science Project. The unsupervised machine learning approach makes the system highly flexible, easily adaptable and provides a domain-independent method of text normalization.

Keywords: SMS, parallel corpus, French, abbreviation/shorthand form normalization, unsupervised learning, character-based machine translation, distributional semantics, Multivec, neural networks, deep learning, word embeddings

1. Introduction

Text messages are a form of computer-mediated communication (CMC). They have raised new discussions in linguistics about their spoken or written usage and their online or offline synchronicity. As the following quote points out, whether text messages are similar to spoken language remains a point of contention: "The degree to which CMC can be associated with speech and/or writing […] often depends on its level of synchronicity" (Van Compernolle 2010: 448).

https://doi.org/10.1075/scl.90.17gho
EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use
© 2019 John Benjamins Publishing Company

Koch and Oesterreicher (2001: 3) have also mentioned two probable factors, namely concept and medium, through the interaction of which a genre can be considered “oral” or “written”. To summarize, we can expect variability from standard French if the language is used in close, online, informal and speech-like communication.

In our work, we approach the issue of normalizing shorthand forms in Belgian French text messages (Fairon 2007) using two approaches. In the first approach, we see normalization as a machine translation task and implement a statistical character-based translation system to generate character mappings from the non-standard text messages to standard French. In our second approach, we use bilingual word embeddings to capture similar items that can appear in the same contexts for non-standard and standard French. Both approaches require the use of parallel corpora for training purposes.

The chapter is structured as follows. First, we give a brief overview of the endeavors in normalizing non-standard languages. Then, we explain word embeddings and their state-of-the-art applications. Section 3 describes the corpus and preprocessing steps for our experiments. In Section 4, we describe our methodologies, tools and experiment setups. An extensive result analysis is conducted in Section 5, followed by the conclusion and an outlook for our future work in Sections 6 and 7, respectively.

2. Previous work

Beaufort et al. (2010: 770) provide a brief overview of the variability that can occur in text messages, such as phonetic plays (“*mer6*” read *“merci”*, “thanks”), phonetic transcription (“*ki*” read *“qui”*, “who”), consonant skeletons (*“ktv”* read *“que tu veux”*, “as you wish”), misapplied, missing or incorrect separators (*“ojourdhui”* read *“aujourd’hui”*, “today”), etc.[1] The causes of these deviations may lie in the constraints of the medium, which only allows 140 bytes per text message, as well as in the layout of phone keypads.

Three methodologies have been applied to normalize text messages, namely, the hidden Markov model (HMM, aka noisy channel model), automatic speech recognition (ASR), and statistical machine translation (SMT).

Choudhury et al. (2007: 63–70) authored the earliest work on texting language (TL). They formally and linguistically analyzed the nature and type of TL. For their work, they used a word-aligned corpus of 1,000 English text messages (with 20,000 tokens), which was created manually. HMM was applied to decode the TL English

[1] These examples are from the Belgian corpus.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

to standard English. Essentially, HMM maps hidden phonetic information to characters. Their approach worked fairly well for unseen words; however, it could not handle self-looping (e.g. *“sooooo”* for “so”) or transposition (*“aks”* for “ask”). They evaluated their methodology on 1,228 unique tokens with an accuracy of 80%.

A hybrid approach, applying phrase-based SMT and ASR to normalize nonstandard forms in text messages, was used by Kobus et al. (2008). The logic behind their use of ASR was that SMS can be seen as “an alphabetic/syllabic approximation of a phonetic form” (Kobus et al. 2008: 443). Phrase-based SMT was used to generate the mapping from the original form to the standard, while ASR was able to find the most probable transformation from graphemes (original form) then to phonemes to graphemes (standard form). Dictionaries of grapheme-phoneme-mappings were created. This chapter is of importance to our work, because it is an early work that tackled the issue of normalization using the Belgian sms4science[2] (created in 2004) corpus. 3,000 unseen messages were tested against this hybrid approach with a word error rate (WER) of 11%. The authors did not report the number of unique original tokens in the test set and investigated both word (e.g. *“d”* to *“de”*) and phrase (e.g. *“oublié2tdir”* to *“oublié de te dire”*) normalization.

Yvon (2010) built on the approaches from Kobus et al. (2008) using the same samples from the Belgian French corpus, and claimed that weighted finite state machines (FSM) worked better than phrased-based SMT. An additional dictionary of abbreviation-standard forms was used before FSM. Finally, a statistical language model (3-gram) was used to enhance the results of FSM. The system performance in Yvon (2010) did not outperform that in Kobus et al. (2008).

Another work on the Belgian corpus was Beaufort et al. 2010, which performed a 10-fold cross validation. The ultimate goal of this work was to generate character alignments without using the phonetic similarities. Sentences were split based on the following standard: “the longest sequence of characters parsed without meeting the same separator on both sides of the alignment” (Beaufort et al. 2010: 774). The key contribution is finding that simple segmentation by spaces is not adequate to capture all the possible character mappings (e.g. *“j esper”* to *“j’espère”*). As we can see from Examples (1a) and (1b), word segmentation could be improved using their proposed strategy. Square brackets are markers for the segmentation boundaries; underlines show missing elements. The approach worked very well with an average WER of 9.3% reported on a test set that should be similar to that used by Kobus et al. (2008) according to the authors (Beaufort et al. 2010: 777).

(1) a. [J esper_] [k__tu] [va_]
b. [J’espère] [que tu] [vas] (Beaufort et al. 2010: 774)

[2] For more information, see <http://www.sms4science.org/?q=en> (13 May 2017).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Pennell and Liu (2011) looked at a character-based SMT approach for normalization of abbreviations in text messages. They proposed a two-step method to tackle this issue. The first step includes performing a character-based SMT (translation and language models with 5-grams) to get translation candidates. The next step used a character-based SMT with contextual information to choose the best translation candidate. They observed that by using context words they achieved a better accuracy score. This work is of particular importance as we used five of the seven types of abbreviations described in this work: deletions, substitutions, repetitions, swaps and insertions. Table 1 gives an overview of the categories of shorthand forms as Pennell and Liu (2011: 978) defines; the examples in Table 1 are selected from the Belgian French corpus.

Table 1. Categories and examples of five shorthand forms

Category	Normalized	Original
Deletions	bibliothèque	biblio
Substitutions	quelle	kel
Repetitions	max	maxxxxx
Swaps	nuit	niut
Insertions	ok	oki

On a data set of tweets (4,661 tweets with 7,769 tokens), the best result achieved was a 69.7% accuracy using contextual information. The authors also tested their methodologies with the same data set as Choudhury et al. (2007) and achieved an accuracy of 60.39%.

Li and Liu (2012) compared a two-step process using phonetic sequences with character-based SMT. In the two-step method, they translated non-standard texts into sequence of phonetic symbols in the International Phonetic Alphabet (IPA). Then the phonetic sequences were transformed into words with the help of a dictionary (mapping of phonemes and graphemes). Finally, they combined the aforementioned methodologies to choose the best translation candidate based on the probabilities. They observed that if the translation probability for the character-based SMT is low, then the candidates given by the two-step process with phonetic sequences worked better. They evaluated their methodologies using the same data set as Choudhury et al. (2007) and concluded that their two-step approach with the phonetic sequences yielded an accuracy of 71.96%.

De Clercq et al. (2013) implemented a simple approach using user-generated contents of different types such as text messages, message board posts, and tweets. They simply performed a phrased-based translation of the corpus. The pairs from phrase table were then fed into a character-based translation system. They draw

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

the conclusion that this combined method achieved the best WER (13% with unigram translation with text messages) than the aforementioned methods employed separately.

Distributional semantics are popular methods to represent words by the context surrounding them. An inspiring work by Sridhar (2015) highlights how effective word embeddings can be employed to learn normalization lexicon.

Distributional semantics originate from Firth's (1957: 11) statement: "You shall know a word by the company it keeps". It means that words appear in the similar contexts have similar meanings. In Examples (2a) to (2d), "Berlin" and "Madrid" are both cities, while "Germany" and "Spain" are both countries. Thus, in a simplified manner, we demonstrate that names of countries and names of cities appear in similar contexts.

(2) a. Berlin is a big city.
b. Madrid is a big city.
c. Germany is a big country.
d. Spain is a big country.

The basic idea of word embeddings is to represent a word as vectors with real numbers in a vector space. There are many ways of creating vectors, amongst which the simplest method is the bag-of-words representation. As shown in Examples (2a) to (2d), our corpus consists of four sentences. The vocabulary size is nine, as we have nine unique tokens, that is "Berlin", "Madrid", "Germany", "Spain", "is", "a", "big", "city", "country". The vector space can be constructed with each unique token as a dimension. The order of word dimensions is defined arbitrary but should be kept consistent for all the vector representations under the same vector space. Thus, the vector representations for "Berlin" and "Madrid" using one-hot encoding (more see Rong 2014: 2) are: [1, 0, 0, 0, 0, 0, 0, 0, 0] and [0, 1, 0, 0, 0, 0, 0, 0, 0, 0], respectively (see Table 1). Using this logic, the maximum length of vectors is equal to the number of unique words in the vocabulary. For a more extensive introduction of computational distributional semantics, see Chapter 20: Computational Lexical Semantics in Jurafsky and Martin (2014: 649–694).

Table 2. Examples of one-hot encoded vectors

	Berlin	**Madrid**	**Germany**	**Spain**	**is**	**a**	**big**	**city**	**country**
"Berlin"	1	0	0	0	0	0	0	0	0
"Madrid"	0	1	0	0	0	0	0	0	0

Word embeddings allow us to perform vector-based calculations. The similarity of words can be computed by the distance between two vectors in the same vector space. Common measure in distributional semantics is the cosine similarity

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

score, which ranges from -1 to 1. The larger the cosine similarity score is, the more similar the vectors are. For a given word embedding, we can calculate the nearest neighbours to that word by ranking their cosine similarity score in relation to that word. The top 1 candidate is usually taken as the best possible candidate.

Some other concepts to understand *word2vec* (Mikolov et al. 2013a, 2013b) are continuous bag-of-words (CBOW) and skip-gram. For CBOW, one tries to predict the target word based on its context words. In Example (3), if target word is “run”, given a context window of five, the context words are two words on the left and right of the target words, namely, “cats”, “usually”, “faster”, “than”. We move the context window alongside the whole sentence, so in the next iteration, the target word will become “faster”, with the context words “usually”, “run”, “than”, “humans”. Skip-gram is the opposite of CBOW, namely, that it tries to predict the context of a word. In our example, given a word “run”, skip-gram predicts the context “cats”, “usually”, “faster”, “than”.

(3) Cats usually run faster than humans.

word2vec is a computationally efficient way of calculating word embeddings using CBOW and skip-gram models. It utilizes negative sampling which is a way of randomly sampling co-occurrence in a corpus. For the co-occurrence with the word “cats” in our imaginary corpus, instead of extracting all the words that co-occur with “cats”, we only sample a small amount of words, for example “always”, “often”, “sometimes”, etc. Negative sampling increases the computational efficiency to calculate word embeddings. As mentioned before, the length of word vectors is usually the length of vocabulary and these vectors are usually very sparse. In order to have a condensed representation of vectors, dimensionality reduction is necessary.

word2vec models implement neural networks with only one hidden layer (more see Rong 2014); therefore, a dimensionality reduction procedure takes place. The final embeddings of a given word is the row vector of the weight matrix between the input layer and the hidden layer.

multivec (Bérard et al. 2016) is an extension of w*ord2vec*. Whereas w*ord2vec* calculates the word vector representation in a single language, *multivec* maps bilingual representations across languages by calculating embeddings for a word in the source language as well as similar word in the same context in the target language. For this reason, *multivec* requires perfectly aligned parallel sentences as input.

Based on what we have discussed before, words with similar semantic meanings tend to appear in similar contexts (e.g. similar vector representations), we shall map the bilingual word embeddings for a small parallel corpus (Example (4)) as shown in Figure 1.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

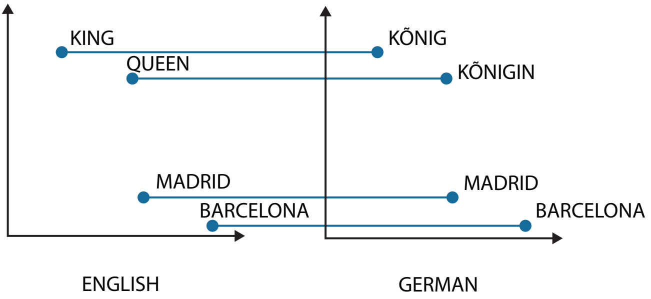


Figure 1. Examples for bilingual word embeddings mapping of English and German

(4) a. The king and queen are coming.
b. Der König und die Königin kommen.
c. Madrid and Barcelona are cities.
d. Madrid und Barcelona sind Städte.

The graph on the left represents the word embeddings in the two-dimensional plane for the English words “king”, “queen”, “Madrid”, “Barcelona”, the graph on the right for their counterparts in German (see Figure 1). The lines represent the cross-linguistic mappings between the words that share the similar contexts.

3. Corpus and preprocessing

3.1 Corpus

We used the Belgian French data from the sms4science project, the goal of which was to carry out linguistic analysis on communication with SMS. The corpus is available as a manually normalized *Excel* file. The file was structured so each row corresponds to a single SMS, and the two columns we selected for our analysis.[3] Each text message in the corpus is regarded a document. The text messages were previously redacted by the creators of the corpus. All the sensitive data were removed, such as telephone numbers, email addresses, etc. We do not have access to the guidelines for normalization. However, based on manual evaluation, we were able to deduce the annotators used different strategies to normalize the texts. As a result, the quality of normalization is inconsistent; hence, it took a lengthy processing step to acquire perfect sentence alignments from the documents. The numbers of tokens and unique spelling variants are listed in Table 3.

[3] The names of the columns we selected are *traduction_normalizee_sans_tag, message_non_normalize* where the rows correspond to the normalized and original versions, respectively.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 3. Tokens and spelling variants in the Belgian French corpus

	Original	Normalized
TOKEN	657,572	681,866
SPELLING VARIANT	46,413	26,687

3.2 Preprocessing

Based on the non-standardized nature of the data, we will see that it will have an effect on the preprocessing and quality of the results. The input for machine translation in *moses* (Koehn et al. 2007) and *multivec* systems (we refer to our two approaches as *moses* and *multivec* hereafter) requires perfectly aligned parallel sentences. In order to generate sentence mapping, we had to identify the sentence boundaries from text messages. A single text message can contain multiple sentences.

To start with, we removed the markups in the corpus for sensitive information (e.g. {???, EMAIL}, {???, NOM}), so that they do not influence the sentence segmenter and work tokenizer. Then we applied the Sentence Segmenter (*nltk.tokenize.sent_tokenize()*) from Natural Language Processing Toolkit (NLTK in Python 2.7, Bird et al. 2009) on our corpus. We found some discrepancies of sentence alignments mainly because punctuations in the original texts can appear within the sentence as emoticons, abbreviations, emphasis, typos, etc.

(5) [OK..pour.20h30-21h,.il.y.aura.2.petit.en.plus.3.et.6.ans,.je.vais.faire.un.ciné.(22h30.][)et.retour.av.1H.du.matin,.je.vous.donne.forfait.40.€?.merci.de.me.repondre.av.midi]

(6) [OK pour 20h30-21h, il y aura 2 petits en plus 3 et 6 ans, je vais faire un cinéma (22h30) et retour avant 1 heure du matin, je vous donne forfait 40 €?][Merci de me répondre avant midi]

As shown in Examples (5) and (6) for one pair of original and normalized text messages, due to the irregularities in the original texts, the segmenter, trained on standard French, was not able to identify the sentence boundaries in an adequate manner. The square brackets in Examples (5) and (6) show the sentence boundaries identified by the segmenter. We decided to manually correct the sentence boundaries. In the end, we obtained 94,982 sentences (aligned in the original and normalized forms) out of 30,000 SMS.

For character-based machine translation, we further split the tokens by word boundaries, then each token was split into individual characters (e.g. *"2 m 1"* to *"d e m a i n"*). Character alignment was performed with GIZA + + (Och & Ney 2003) in *moses*. We trained the machine translation system with the sentences of the same length, which takes 73.5% of all the materials from the corpus. The numbers

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

of tokens and unique spelling variants are listed in Table 3. With *multivec*, the inputs are pairs of aligned parallel sentences. The numbers of tokens and spelling variants can be found in Table 4.

Table 4. Tokens and spelling variants for *moses* training and testing

	Training		**Test**	
	Original	**Normalized**	**Original**	**Normalized**
TOKEN	83,926	83,926	10,490	10,490
TYPE	13,563	9,069	3,360	2,439

4. Methodologies, tools and experiments

4.1 Methodologies

As we discussed in the literature review, we can use the original and normalized text messages as parallel sentences in machine translation systems when we consider the original text message as source language and the normalized text message as target language. Since the text messages have monotonous word alignments, given a reasonable word segmentation strategy, character-based SMT can be applied to generate the character mappings between the two versions of text messages.

Using the logic that similar words tend to appear in similar contexts, we can use *multivec* as a method of translating from the original texts to the standardized texts, because the shorthand forms bear the same context, even if they appear in different spelling variants.

4.2 Tools and experiments

multivec

For *multivec* we use the entire corpus for training the model (corpus size see Table 3). The unit of analysis in *multivec* is the word. The setup of *multivec* is as follows:

1. Dimensions in the vectors: 100
2. Context windows size: 5
3. Minimum absolute counts in corpus: 5
4. Learning rate: 0.05
5. Iterations: 5
6. Subsampling: 0.001
7. Negative sampling: 5
8. The remaining parameters: default

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

We observe that true casing works better than lower casing. Moreover, we achieve the following accuracy scores in the nearest neighbour list for the top 1 item and top 3 items: 55.3%, 59.0%.

moses
The unit of analysis in *moses* is the character. In the preprocessing phase we took only sentences with the same length, because the simple segmentation by word boundaries can lead to mismatch in the original and normalized sentences (for corpus size see Table 4). We tried language models of up to 7-grams and found out that the normalization accuracy of shorthand forms stabilized at 7-grams.

The best result is given by taking the top 1 candidate in the 10-best list, language model 7-gram, grow-diag-final-and. The average length of a token in our corpus is 7 (6.76 for *moses*, 5.7 for *multivec*), which might explain why the 7-gram language model scored the best.

5. Results analysis

We took a test set of 10,490 shorthand forms for both *moses* and *multivec*. We looked at the recall of the two approaches and their overlaps in identifying shorthand forms. To evaluate accuracy, we took the overlapping part as the test set. *multivec* retrieved 5,809 instances, whereas *moses* found 5,283. We report only accuracy of the task, because the precision of retrieval is 100%. The accuracy score is equal to the recall of shorthand forms in the test set. The two approaches overlap in 3,987 retrieved instances, with discrepancies of 1,821 for *multivec* and 1,291 for *moses*, respectively. Spelling variants of a lemma refer to the variability of that lemma in different forms. For the lemma *"demain"*, there are plenty of spelling variants in our corpus such as *"dm1"*, *"dm 1"*, *"dem1"*, *"dem 1"*, *"2m1"*, *"2main"*, *"2min"*, etc.

We analyzed the retrieval discrepancies extensively at the token and spelling variant levels in the following sections. Because we trained the systems of *multivec* and *moses* with different sizes of training materials, normalization of retrieved instances against the size of training materials is needed in order to evaluate the output on the same basis (Table 5). As seen in Figure 2 and Table 5, *moses* recognizes a lot more cases than *multivec* on both the token and spelling variant levels. However, in the test set, we notice that *multivec* is able to correctly identify more tokens, while *moses* is better at identifying spelling variants (see Figure 2).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 5. Normalized retrieved rate for tokens and spelling variants

	moses	multivec
TOKEN (retrieved)	1,291	1,821
SPELLING VARIANT (retrieved)	903	473
TOKEN (training set)	83,926	657,572
SPELLING VARIANT (training set)	13,569	46,413
NORMALIZED TOKEN%	1.50%	0.28%
NORMALIZED SPELLING VARIANT%	6.65%	1.02%

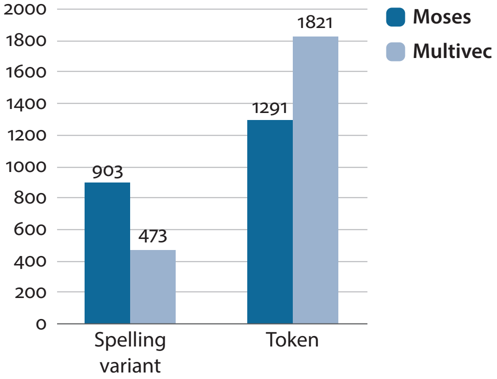


Figure 2. Discrepancies of tokens and spelling variants

When we compare *moses* with *multivec*, *multivec* recognizes the most deletions (see Figures 3 and 4). Although *moses* has less training materials, it is better at recognizing substitutions and repetitions. There are not many cases of swaps and insertions in the test set; thus, it is difficult to judge the performances of the two systems on these categories.

We looked at the different categories of shorthand forms recognized in the test set and observed that *moses* and *multivec* have different preferences for normalizing different categories. Deletions are the most common category in the test set, followed by substitutions. The number of deletions is 3 times that of substitutions. The number of substitutions is 20 times that of repetitions. Swaps and insertions are much less observable in the test set (see Figures 3 and 4).

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

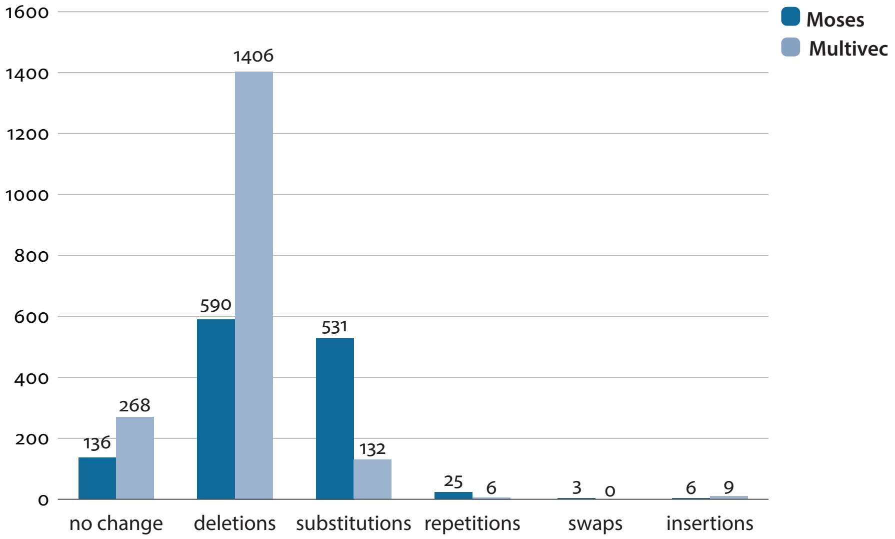


Figure 3. Accurate normalization of token level for *moses* and *multivec* in five categories

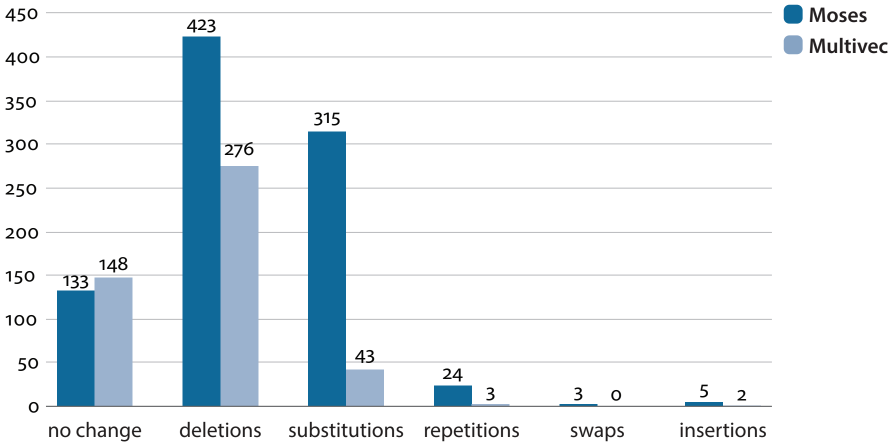


Figure 4. Accurate normalization of spelling variant level for *moses* and *multivec* in five categories

As demonstrated in Table 6, *multivec* is able to find severely shorted forms (“we” for “weekend”, “kel” for “quelle”). *Moses* is capable of finding highly varied forms of spelling variants.

When we compare *moses* with *multivec*, we see that the most common type that has been recognized by *multivec* is deletions. Although *moses* has fewer training materials, it is better at recognizing substitutions and repetitions. There are not many cases of swaps and insertions in the test set; hence, it is difficult to judge the performances of the two systems on these categories.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 6. Examples from *moses* and *multivec* for five shorthand categories

	moses	**multivec**
Deletions	“ojrd8” (“aujourd’hui”)	“we” (“weekend”)
Substitutions	“ki” (“qui”)	“kel” (“quelle”)
Repetitions	“groos” (“gros”)	“bizzz” (“biz”)
Swaps	“niut” (“nuit”)	–
Insertions	“dorlotter” (“dorloter”)	“oki” (“ok”)

*normalized word in bracket

When it comes to the spelling variants, we see that across all the categories, *moses* captures most of the normalization cases accurately. In the test set, it highly outperforms *multivec* when it comes to substitutions. In general, *moses* has a higher rate of recall on different types, whereas *multivec* has high precision, as it recognizes words that appear in similar contexts all the time.

Some other interesting observations are that *moses* is flexible on different spelling variants (see Table 7) and is good with temporal expressions (see Table 8). As we see in Table 7, for the normalization form *“aujourd’hui”,* nine different spelling variants were found in the test set, including the lemma. Both systems have different preferences in “translating” from the original to the normalized forms.

We noticed that *multivec* recognizes always the same items, for example *“auj”* (7 times), *“ajd”* (4 times). *Moses* recognizes six different spelling variants, and except for *“aujourdhui”* (3 times), the other five spelling variants are hapax legomena. *“auj”* and *“ajd”* are extremely difficult to for *moses* to normalize, as long-distant deletion of character sequences is involved.

We also noticed that *moses* is good at finding temporal expressions (see Table 8). In these forms, the variation lies in separators between hours and

Table 7. Examples for “*aujourd’hui*” in different spelling forms

Original	**Normalized**	**Approach**	**Frequency**
auj	aujourd’hui	*multivec*	7
adj			4
aujourd’hui			1
aujourdhui		*moses*	3
aujourdwi			1
ojourdhui			1
aujourdui			1
ojrd8			1
ojordui			1

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 8. Examples for temporal expressions

Original	Normalized
11h58	11h58
12:30	12h30
17.40h	17h40
18.30	18h30
19	19h
19h3o	19h30
20.49	20h49
8.30	8h30

minutes, for example *“h”*, *“:”*, *“.”*. *Moses* is able to normalize extremely varied form such as *“17.40h”* to its normalized form *“17h40”*.

Moses is highly flexible with out-of-vocabulary words (words that have not been “seen” in its training materials) and is good at identifying different spelling variants. During the training of *multivec*, it omitted out-of-vocabulary words. Nonetheless, *moses* has the advantages of treating every normalization as a translation task and tries to find the most probable character mappings from the original form to the normalized form.

However, *Moses* can bring in noises in character mapping, because it calculates the path with the highest joint probabilities. For example, the word *“l’université”* was wrongly mapped to *“l’universitait”* by taking the path with the highest probability. The probabilities of character mapping are taken from the phrase table and path 1 is the most probable path amongst all the mapping strategies. Path 2 is an example of another possible way to normalize the word *“l’université”*, and since it has a much lower probability compared to path 1, *moses* opted for path 1 (see Table 9).

Table 9. *Moses* phrase table and the different paths of normalization

Path 1	l’uni → l’uni (0.75)	v → v (0.86)	s → s (0.71)	i → i (0.86)	té →tait (0.66)	l’universitait (0.26)
Path 2	l’uni → l’uni (0.75)	ver → ver (0.28)	si → si (0.6)	té →tait (0.66)		l’université (0.08)

To summarize, using *multivec* means looking for existing items in a dictionary. If the item is unlisted in the dictionary, *multivec* cannot solve the normalization tasks. On the contrary, *moses* treats its normalization task by applying character mapping rules to the non-standard forms. It attempts to come up with the most plausible path.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

6. Conclusion

Using *multivec* to normalize shorthand forms has its advantages as this process is always consistent and runs extremely fast. Moreover, it is able to retain contextual information; thus, highly abbreviated items can be deciphered by *multivec*. Unfortunately, this method is highly intolerant of unseen items in training materials. Consequently, it requires large input amounts of parallel materials.

Moses is highly flexible when it comes to translations of out-of-vocabulary items. Moreover, it is able to retrieve multiple spelling variants. As discussed before, *moses* chooses the path of the highest joint probability. Thus, it can occur that during the character mapping process, noises (an incorrect mapping with a higher probability) are inserted into the output. Furthermore, *moses* requires a long training time and we have to train the language model accordingly, that is we have to run different models with different parameters until reaching a satisfactory result, which can lead to even longer training and tuning time.

As we do not have access to the evaluation sets in the previous studies on the Belgian French corpus (see Kobus et al. 2008; Yvon 2010 and Beaufort et al. 2010) and we only investigated word normalization in our experiments, we cannot compare our results directly with those in the state-of-the-art study (i.e. Beaufort et al. 2010). Our contributions are as follows:

1. To the best of our knowledge, we are the first study that applied the character-based SMT to normalize the Belgian French text messages.
2. We applied word embeddings that encode contextual information to increase recall and precision of normalization.
3. We identified the efficacy and necessity of combining the two approaches in normalization: character-based SMT (*moses*) and word embeddings (*multivec*).
4. Our detailed result analyses provide further insights into future work.

7. Future work

As we did not include contextual information into *moses*, as it could be quite challenging to encode contextual information in the character system, which requires precise segmentation of word boundaries. We suggest that in the future, we could combine the two approaches as the word embeddings from *multivec* already entails the contextual information of a word. It is plausible to identify the candidates for character-based SMT by *multivec*, and then use the parallel pairs as input for machine translation systems.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Another point where we could improve is incorporating parallel sentences of different lengths, aka phrase normalization, where one side of the pair is a multi-word expression (e.g. *"ktv"* to *"que tu veux"*). We also envisage using more materials from the sms4science project, namely the French sms4science corpus collected in France. Finally, in the realm of embeddings, we would like to extend our methodologies of creating embeddings to units on the sub-word level. *fastText* (Bojanowski et al. 2016) provides the theoretical framework for applying this methodology to any language. We assume that mapping the original and normalized texts on the level of sub-word unit can reduce the number of out-of-vocabulary items, as well as omit noises as in *moses*. Moreover, as the sub-word embeddings entail morphological and contextual information the mapping quality could be improved. We will address the aforementioned open issues in our future experiments.

Acknowledgment

Both authors contributed equally to the project. Authors are listed in alphabetical order. The authors appreciate the valuable comments and support provided by Prof. Dr. Martin Volk and Natalia Korchagina from the Institute of Computational Linguistics at the University of Zurich and thank Heath Gordon and the anonymous reviewers for proofreading. All remaining errors are ours.

References

Beaufort, Richard, Roekhaut, Sophie, Cougnon, Louise-Amélie & Fairon, Cédrick. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS Messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 770–779.

Bérard, Alexandre, Servan, Christophe, Pietquin, Olivier & Besacier, Laurent. 2016. Multivec: A multilingual and multilevel representation learning toolkit for NLP. *The 10th edition of the Language Resources and Evaluation Conference*, 4188–4192.

Bird, Steven, Loper, Edward & Klein, Ewan. 2009. *Natural Language Processing with Python*. San Francisco CA: O'Reilly Media.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand & Mikolov, Tomas. 2016. *Enriching Word Vectors with Subword Information*. <https://arxiv.org/abs/1607.04606> (13 May 2017).

Choudhury, Monojit, Saraf, Rahul, Jain, Vijit, Sudeshna, Sarkar & Basu, Anupam. 2007. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition* 10(3): 157–174. https://doi.org/10.1007/s10032-007-0054-0

De Clercq Orphée, Schulz, Sarah, Desmet, Bart, Lefever, Els, Hoste, Véronique. 2013. Normalization of Dutch user-generated content. *Proceedings of 9th International Conference on Recent Advances in Natural Language Processing*, 179–188. Berlin: Springer.

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Fairon, Cécrick, Klein, Jean R. & Paumier, Sébastien. 2007. *Le langage SMS: étude d'un corpus informatisé à partir de l'enquête 'Faites don de vos SMS à la science'*. Louvain-la-Neuve: Presses universitaires de Louvain.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis*, 1–32. Oxford: Blackwell.

Jurafsky, Daniel & Martin, James H. 2014. *Speech and Language Processing*. Englewood Cliffs NJ: Prentice Hall.

Kobus, Catherine, Yvon, François & Damnati, Géraldine. 2008. Normalizing SMS: Are two metaphors better than one? *Proceedings of the 22nd International Conference on Computational Linguistics* 1, 441–448.

Koch, Peter & Oesterreicher, Wulf. 2001. *Gesprochene und geschriebene Sprache. Französisch, Italienisch, Spanisch*. Berlin: De Gruyter.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Brooke Cowan, Nicola, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra & Herbst, Evan. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. https://doi.org/10.3115/1557769.1557821

Li, Chen & Liu, Yang. 2012. Normalization of text messages using character-& phone-based machine translation approaches. *Proceedings of 13th Annual Conference of the International Speech Communication Association*, 2330–2333.

Mikolov, Tomas, Chen, Kai, Corrado, Greg & Dean, Jeffrey. 2013a. Efficient estimation of word representations in vector space. *The Workshop Proceedings of the International Conference on Learning Representations*. <https://arxiv.org/abs/1301.3781> (13 May 2017).

Mikolov, Thomas, Ilya, Sutskever, Chen, Kai, Corrado, Greg & Dean, Jeffrey. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*.<https://arxiv.org/pdf/1310.4546.pdf> (13 May 2017).

Och, Franz Josef & Ney, Hermann. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1): 19–51. https://doi.org/10.1162/089120103321337421

Pennell, Deana L. & Liu, Yang. 2011. A character-level machine translation approach for normalization of SMS Abbreviations. *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP)*: 974–982.

Rong, Xin. 2014. word2vec parameter learning explained. <https://arxiv.org/abs/1411.2738> (13 May 2017).

sms4science project. 2004. <http://www.sms4science.org/?q=en> (13 May 2017).

Sridhar, V. K. R. 2015. Unsupervised text normalization using distributed representations of words and phrases. *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*: 8–16.

Van Compernolle, Rémi A. 2010. The (slightly more) productive use of *ne* in Montreal French chat. *Language Sciences* 32(4): 447–463. https://doi.org/10.1016/j.langsci.2009.07.004

Yvon, François. 2010. Rewriting the orthography of SMS messages. *Natural Language Engineering* 16(2): 133–159. https://doi.org/10.1017/S1351324909990258

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

Subject index

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use

This volume assesses the state of the art of parallel corpus research as a whole, reporting on advances in both recent developments of parallel corpora – with some particular references to comparable corpora as well – and in ways of exploiting them for a variety of purposes. The first part of the book is devoted to new roles that parallel corpora can and should assume in translation studies and in contrastive linguistics, to the usefulness and usability of parallel corpora, and to advances in parallel corpus alignment, annotation and retrieval. There follows an up-to-date presentation of a number of parallel corpus projects currently being carried out in Europe, some of them multimodal, with certain chapters illustrating case studies developed on the basis of the corpora at hand. In most of these chapters, attention is paid to specific technical issues of corpus building. The third part of the book reflects on specific applications and on the creation of bilingual resources from parallel corpora. This volume will be welcomed by scholars, postgraduate and PhD students in the fields of contrastive linguistics, translation studies, lexicography, language teaching and learning, machine translation, and natural language processing.

ISBN 978 90 272 0234 5

9 789027 202345

JOHN BENJAMINS PUBLISHING COMPANY

EBSCOhost - printed on 2/10/2023 1:23 AM via . All use subject to https://www.ebsco.com/terms-of-use