The use of corpora in legal and institutional translation studies

Directions and applications

Fernando Prieto Ramos
University of Geneva

Research in legal and institutional translation within the realm of Legal Translation Studies (LTS) has greatly benefited from embracing the advances of Corpus Linguistics in the past few decades. This paper provides an overview of corpus-based approaches in LTS and illustrates their increasing prominence and sophistication through the description of seven selected representative projects, including a wide range of corpus types, translation contexts, legal genres, jurisdictions, sizes and languages. The comparative examination of these studies confirms the relevance of corpus methods for LTS, the need to integrate quantitative and qualitative considerations (crucially including legal parameters) into corpus-building criteria, as well as the correlation between research scope and methodological nuance in ensuring corpus suitability.

Keywords: legal corpora, Corpus-Based Legal Translation Studies (CBLTS), research methodology, parallel corpora, comparable corpora, representativeness, legal translation, institutional translation

Introduction: Corpus-based legal translation studies

Advances in Corpus Linguistics have become increasingly popular in Translation Studies since the mid-1990s (see e.g. Baker 1993, 1995; Laviosa 2002; Olohan 2004; Zanettin 2012), leading to what is known as Corpus-Based Translation Studies. This trend, which has been facilitated by software developments and the digitalization of large volumes of text, can be associated with the "technological turn" in Translation Studies (Cronin 2010) and a pronounced shift towards "datafication" and empiricism (Snell-Hornby 2006, 114). Research into legal and institutional translation, and legal discourses more broadly, has not been an exception. The interdisciplinary field of Legal Translation Studies (LTS) has, in the same period,

https://doi.org/10.1075/ts.00010.pri
Translation Spaces 8:1 (2019), pp. 1–11. issn 2211-3711 | e-issn 2211-372x

witnessed a marked expansion characterized by growing scope, thematic diversification and methodological sophistication (Biel 2017; Prieto Ramos 2014a). As noted by Biel and Engberg (2013, 5), corpus-based methodologies stand out in this context as the "most frequently represented" approach.

Given the focus of translation on text, it is no surprise that research in this field has benefited from the development of new tools to analyze discourse features and patterns. Already in the early 2000s, Bhatia et al. (2004, 203) described corpus-based studies as "so popular that one rarely finds a textual study without the use of computerized corpora." What has clearly evolved since then, apart from the popularity of these approaches, is their degree of rigor and detail, enriching analyses of translation issues with more empirical data.

Nowadays, in legal and institutional translation, as in other areas of translation, corpora are not only designed for research, but are often employed to support translation practice and training as well (see e.g. Biel 2018; Monzó Nebot 2008; Pontrandolfo 2012). Corpus-based approaches are yielding new insights into legal discourses and translation practices in multiple languages and jurisdictions. While *comparable corpora* (i.e. sets of monolingual texts compiled according to the same criteria) are frequently used to describe cross-linguistic variation and translated text features as opposed to non-translated text conventions (e.g. Biel 2014; Mori 2018; Vanden Bulcke 2013), *parallel corpora* (i.e. sets of source texts and their translations) prevail in the analysis of translation decision-making and quality (e.g. Prieto Ramos and Guzmán 2018; Simonnæs and Whittaker 2013; Trklja and McAuliffe 2018). *Mixed corpus design*, combining both types of corpora, may also be relevant to integrate the examination of source texts into the cross-linguistic analysis of patterns (see e.g. Biel 2016). Despite the proliferation of corpus-based studies in the field and the growing availability of resources developed by public institutions, a review of legal corpora reveals several persistent gaps, namely:

- Parallel corpora are significantly scarcer, particularly multilingual corpora (see e.g. Pontrandolfo 2012, 133), for which corpus analysis software is underdeveloped (see e.g. Cerutti 2017 for a recent analysis of concordancers for a trilingual corpus in the framework of the LETRINT project).[1]
- Corpora for inter-systemic legal translation are uncommon, especially in the case of lesser-used languages, which is in line with the traditional preeminence

[1] "Legal Translation in International Institutional Settings: Scope, Strategies and Quality Markers", led by Fernando Prieto Ramos and supported by the Swiss National Science Foundation through a Consolidator Grant.

of major European languages in LTS (see Biel 2018, 34; Biel and Engberg 2013, 2).
- The traditional focus on legislative genres in LTS (see Biel 2016, 199; Prieto Ramos 2014a, 265), together with the difficulty of accessing private law documents, have resulted in fewer accessible corpora of non-legislative genres.

In an attempt to partially fill these gaps and illustrate the dynamism and usefulness of corpus-based approaches in LTS, seven research projects have been selected to form this specialized volume on Corpus-Based LTS (or CBLTS), including corpus-based and corpus-driven studies on legal and institutional translation.[2] These projects[3] are considered representative of the development of LTS beyond legislative translation, including: bilingual inter-systemic translation and international multilingual translation contexts, public and private law documents, less explored languages in LTS such as Portuguese and Thai, and authors from Europe, Latin America and Asia. The advantages and challenges of using different types of corpora (comparable or parallel; monolingual, bilingual or multilingual) and analytical methods are explicitly addressed by these authors in the light of their research aims.

We will describe the main features of the relevant corpora focusing on methodological issues of corpus compilation and analysis, with a view to identifying commonalities and potential correlations between research goals and quantitative and qualitative features of corpora in this field. Corpora will therefore be regarded here not only as instrumental to research, but also as an important object of study itself, which reflects the long-debated question of the disciplinary status of Corpus Linguistics (see e.g. McEnery et al. 2006, 7–8). In fact, the instrumentality of an area of study to adjacent fields of research is not alien to LTS, as comparative legal analysis also plays a key instrumental role in legal translation.

[2] In this paper, by analogy with “Corpus-Based Translation Studies,” references to “corpus-based” studies will be used as an overarching denomination including “corpus-driven” approaches (see distinction e.g. in Tognini-Bonelli 2001, 84: respectively, using a selection of examples “to support linguistic argument or to validate a theoretical statement,” as opposed to fully adapting statements to “the evidence provided by the corpus”).

[3] The majority were presented at the Transius International Conference on Legal and Institutional Translation held in Geneva in June 2018 (http://transius.unige.ch/en/conferences-and-seminars/tc18/cfp/).

Tailoring corpus design to research goals: A comparative overview

The first four studies of the volume use corpora or sub-corpora of legal texts of EU or international jurisdictions (including monolingual and parallel corpora), while the remaining three focus on corpora at national level, either for inter-systemic (UK-Portugal and USA-Peru comparable corpora) or intra-systemic translation (Swiss parallel corpus). The main features of all these corpora are summarized in Table 1, and will be compared in this Section, including research purposes, genre, translation context, languages and jurisdictions involved, time span, size and sources.

In the first part, *Katia Peruzzo*'s study employs a monolingual corpus composed of 16 texts (*ECtHR-IT*). The design of this corpus responds to the specific aims of her research in developing a methodology for extracting loan words that refer to Italian legal concepts and institutions in judgments delivered by the European Court of Human Rights (ECtHR) in English, and analyzing the techniques applied to convey such system-bound concepts in the target language. Drawing on the classification of legal terminology in international institutional settings proposed by Prieto Ramos (2014b), the author illustrates how national legal concepts are not infrequent in these settings and often call for the combination of translation techniques rather than a stand-alone borrowing of the original term.

Mali Satthachai and Dorothy Kenny focus on another key area of legal translation: legislative genres. This study, the first of its kind on English-Thai translation in LTS, delves into the distinctive features of translation of deontic modality into Thai by presenting a comparative analysis of a parallel corpus of international treaties translated from English and a monolingual corpus of original Thai legislative texts (*Thai-LEG*). Following Biel's (2014) concepts of equivalence and textual fit, their inter-linguistic and intra-linguistic comparisons reveal, among other findings, that modal verbs are overrepresented in translation into Thai compared with non-translated texts.

In a similar vein but at a larger scale, *Łucja Biel, Dariusz Koźbiał and Katarzyna Wasilewska* explore the formulaicity (understood as high-frequency multi-word sequences according to Biber and Barbieri 2007) of EU translations into Polish in four institutional genres (legislation, judgments, reports and websites). They analyze several parallel corpora of EU translated texts and comparable Polish monolingual corpora from national sources compiled as part of the *Polish EUROLECT* project (Biel 2016). The study confirms a significant correlation between formulaicity and genres, a higher formulaicity of EU Polish translations (with formulaic profiles of their own) as opposed to non-translations, and a limited overlap of bundles between the two.

Table 1. Main features of selected corpora

Authors and research aims	Corpus name * and type	Genre and translation context	Languages and jurisdictions	Time span	Size and sources
Katia Peruzzo: Develop a methodology to extract loan words semi-automatically. Analyze translation techniques.	ECtHR-IT*: one monolingual corpus.	ECtHR judgments (supranational institutional translation).	English (texts with references to original Italian terms). European jurisdiction.	2000–2018	16 texts, 227,393 tokens (HUDOC database).
Mali Satthachai and Dorothy Kenny: Explore legislative translation from English into Thai. Analyze how instances of deontic modality are translated into Thai.	Thai-LEG*: one bilingual parallel corpus, one monolingual corpus.	International and domestic legislative texts (inter-systemic translation from international to national jurisdiction).	English and Thai. International and Thai jurisdictions.	Bilingual corpus: 1950–2018 Monolingual corpus: 1970–2018	173 texts, 1,568,780 tokens (websites of Thai government agencies, Thailand's Office of the State Council).
Łucja Biel, Dariusz Koźbiał and Katarzyna Wasilewska: Explore the formulaicity of EU translations into Polish.	PL EUROLECT: six parallel and comparable bilingual corpora, four monolingual comparable corpora.	Legislation, judgments, reports and websites (supranational institutional translation).	English and Polish. EU and Polish jurisdictions.	2011–2015 (except for websites: 2015–2016)	11,550 texts, 43.7m tokens (EUR-Lex, Wolters Kluwer SA's Lex, InfoCuria, Polish Supreme Court, Public Register of the European Commission, Polish ministries' websites).

* Corpora without a proper name will be given one for ease of reference in this study. They are marked with an asterisk.

Table 1. *(continued)*

Authors and research aims	Corpus name* and type	Genre and translation context	Languages and jurisdictions	Time span	Size and sources
Fernando Prieto Ramos, Giorgina Cerutti and Diego Guzmán: Map institutional translation (LINST). Categorize multilingual institutional texts from a legal perspective; quantify translation per institutional function and genre; define the scope of institutional legal translation (LETRINT 0). Analyze discourse features, translation patterns and quality indicators (LETRINT 1, LETRINT 1+).	LINST: three monolingual comparable corpora. LETRINT 0: three monolingual comparable corpora. LETRINT 1: one parallel trilingual corpus. LETRINT 1+: one parallel trilingual corpus.	Several UN, EU and WTO genres of law-making, implementation monitoring and adjudication (supranational institutional translation).	English, French and Spanish. International and EU jurisdictions.	2005, 2010 and 2015	LINST: 513,640 texts, 1.71 billion tokens. LETRINT 0: 340,979 texts, 1.18 billion tokens. LETRINT 1: 7,918 texts, 25.76m tokens. LETRINT 1+: ongoing compilation. (UN's Official Document System, WTO Documents Online, EUR-Lex, CJEU database, European Council Document Register, European Parliament Public Register of Documents, Register of Commission Documents).
Mary Ann Monteagudo Medina: Explore denominative variation.	INC-US-Per*: two monolingual comparable corpora.	Business incorporation documents (inter-systemic translation).	English and Spanish. US and Peruvian jurisdictions.	2016–2018	104 texts, 76,935 tokens (US and Peruvian company sources)
Tereza Passos e Sousa Marques Afonso and Maria do Céu Henriques de Bastos: Perform a contrastive legal and linguistic analysis to facilitate translation.	PoA-UK-P*: two monolingual comparable corpora.	Powers of attorney (inter-systemic translation).	English and Portuguese. UK and Portuguese jurisdictions.	2000–2018	36 texts, 14,564 tokens (UK and Portuguese private sources)
Annarita Felici and Cornelia Griebel: Evaluate language clarity on the basis of plain language guidelines considering the translation variable.	INSU-CH*: one parallel trilingual corpus.	Official insurance leaflets (intra-systemic translation in trilingual national context).	German, French and Italian. Swiss jurisdiction.	2016–2018	30 texts, 56,609 tokens (Information Centre, OASI/DI website, Switzerland)

* Corpora without a proper name will be given one for ease of reference in this study. They are marked with an asterisk.

With their focus fixed on the methodology of the far-reaching *LETRINT* project (see footnote 1), *Fernando Prieto Ramos, Giorgina Cerutti and Diego Guzmán* describe the challenges of developing multiple sets of comparable and parallel corpora to define the scope of international institutional translation (as exemplified by the EU, the UN and the WTO), and to analyze the discourse features and translation patterns and quality of legal genres in the three common languages of these settings (English, French and Spanish). In order to ensure representativeness of sampling frames of law-making, implementation monitoring and adjudication genres, a multi-layered sequential approach to corpus-building was tailored to the LETRINT research aims, starting with a full mapping and categorization of institutional texts from a legal perspective, followed by an innovative combination of stratified sampling techniques integrating quantitative and qualitative criteria. These criteria are made explicit and traceable in selection records for the sake of transparency and enhanced re-usability.

In the second part of the volume, entirely devoted to national legal settings, *Mary Ann Monteagudo Medina* exploits a corpus of 104 business incorporation documents from the United States and Peru (*INC-US-Per*) to analyze denominative variation for the purposes of English-Spanish translation. After contextualizing the relevant genres in each legal system, the author highlights significant asymmetries, including terminological variants that are specific to certain American states or types of business organization.

In the following paper, *Tereza Passos e Sousa Marques Afonso and Maria do Céu Henriques de Bastos* adopt a similar approach to compare the system-bound characteristics of powers of attorney in England, Wales and Northern Ireland, and their closest corresponding genre in Portugal, *procuração*. More specifically, they outline the legal framework and typical structure of these genres, and then focus on terminological extraction from their two monolingual comparable corpora (*PoA-UK-P*) in order to illustrate the application of translation techniques in cases of conceptual asymmetry.

Finally, *Annarita Felici and Cornelia Griebel* investigate language clarity in a parallel, trilingual and intra-systemic corpus (*INSU-CH*) that comprises the French, German and Italian versions of ten leaflets on old-age and survivor's insurance and disability insurance published by the Swiss authorities. Following plain language guidelines, they measured readability using indices and syntactic tagging, and triangulated the results with a more qualitative analysis of linguistic problems that point to gaps in plain communication. These problems, for which the authors suggest text simplifications, are generally reproduced in all language versions, as translations seem to be conditioned by the complexity of applicable legal requirements and the need to preserve an identical brochure layout.

Concluding remarks

Our comparative analysis of contemporary applications of corpus-based approaches to research in legal and institutional translation studies confirms the tremendous value of these methodologies for meeting research needs in LTS. As in other strands of translation research, the validity of results can be compromised if corpus-building and corpus-querying parameters are not adequately addressed, in particular to ensure the relevance, representativeness and balance of corpus components. In the case of LTS, as illustrated by the projects selected for this special issue, corpus-building parameters must integrate legal considerations and the situational factors that condition legal translation and drafting practices.

More specifically, corpus designers in this area must contextualize specific genres in their jurisdictions and branches of law, and determine their connections through inter- or intra-systemic translation or co-drafting, whether at national or international level. As noted by Koester (2010, 67), familiarity with the context is expected of specialized corpus compilers and analysts in order to balance quantitative and qualitative dimensions. In an interdisciplinary area characterized by a high variability of translation scenarios and a remarkable diversity of communicative settings and genres, this contextualization is as crucial for research design as it is for legal and institutional translation methodologies themselves.

While the above applies to all the corpora examined, the papers collected in this volume also confirm that the challenges of corpus design and compilation are relative to the scope and level of ambition of each research project. The broader the area of investigation and the aspirations for generalization, the more complex (and the riskier) the definition of corpus sampling criteria that will ultimately underpin the acceptability of the research findings. For example, whereas corpora built for the purposes of case studies on specific genre conventions, terminology or translation techniques (ECtHR-IT, INC-US-Per, PoA-UK-P and INSU-CH) include between 16 and 104 texts, and between 14,564 and 227,393 tokens, larger-scale projects on entire settings of institutional translation (PL EUROLECT and LETRINT) required the compilation of multi-genre, multi-million-word corpora, as well as more nuance about genre representativeness and text selection methods. In the latter case, corpus-tailoring even demanded an unprecedented combination of stratified sampling techniques. Finally, mid-way along this spectrum, generalizations on modality in Thai translated legislative texts were supported by the comparative analysis of a parallel corpus and a monolingual corpus (Thai-LEG) of a combined size of more than 1.5 million tokens.

Overall, the above corpora, all fit for their research purposes, show that the value and methodological soundness of specialized corpora are not just a matter of size but also, crucially, of qualitative properties in the light of research aims, and they necessarily entail "a marriage of perfection and pragmatism" (McEnery et al. 2006,73). It is hoped that the methods outlined in this volume will stimulate greater rigor and further innovation in the application of corpus-based approaches for the sake of quality and traceability of research into legal and other specialized translation.

Acknowledgements

I would like to thank all the contributors and the journal editors for their fruitful cooperation, as well as Diego Guzmán for his technical assistance.

References

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies. Implications and Applications." *Text and Technology: In Honour of John Sinclair*, edited by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233–250. Amsterdam and Philadelphia: 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

Bhatia, Vijay Kumar, Nicola Langton, and Jane Lung. 2004. "Legal Discourse: Opportunities and Threats for Corpus Linguistics." *Discourse in the Professions. Perspectives from Corpus Linguistics*, edited by Ulla Connor, and Thomas A. Upton, 203–231. Amsterdam and Philadelphia: John Benjamins. https://doi.org/10.1075/scl.16.09bha

Biber, Douglas, and Federica Barbieri. 2007. "Lexical Bundles in University Spoken and Written Registers." *English for Specific Purposes* 26 (3): 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biel, Łucja. 2014. *Lost in the Eurofog: The Textual Fit of Translated Law*. Frankfurt am Main: Peter Lang. https://doi.org/10.3726/978-3-653-03986-3

Biel, Łucja. 2016. "Mixed Corpus Design for Researching the Eurolect: A Genre-based Comparable-parallel Corpus in the PL EUROLECT Project." *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*, edited by Ewa Gruszczyńska, and Agnieszka Leńko-Szymańska, 197–208. Warsaw: Instytut Lingwistyki Stosowanej.

Biel, Łucja. 2017. "Researching Legal Translation: A Multi-perspective and Mixed-method Framework for Legal Translation." *Revista de Llengua i Dret / Journal of Language and Law* 68: 76–88.

Biel, Łucja. 2018. "Corpora in Institutional Legal Translation: Small Steps and the Big Picture." *Institutional Translation for International Governance: Enhancing Quality in Multilingual Legal Communication*, edited by Fernando Prieto Ramos, 25–36. London: Bloomsbury.

Biel, Łucja, and Jan Engberg. 2013. "Research Models and Methods in Legal Translation." *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 12: 1–11.

Cerutti, Giorgina. 2017. "Evaluating Tools for Legal Translation Research Needs: The Case of Fourth-generation Concordancers." *Legal Translation and Court Interpreting: Ethical Values, Quality, Competence Training*, edited by Annikki Liimatainen, Arja Nurmi, Marja Kivilehto, Leena Salmi, Anu Viljanmaa, and Melissa Wallace, 357–391. Berlin: Frank and Timme.

Cronin, Michael. 2010. "The Translation Crowd." *Revista Tradumàtica* 8. Accessed 18 December, 2018. http://www.fti.uab.es/tradumatica/revista/num8/articles/04/04central.htm. https://doi.org/10.5565/rev/tradumatica.100

Koester, Almut. 2010. "Building Small Specialised Corpora." *The Routledge Handbook of Corpus Linguistics*, edited by Michael McCarthy, and Anne O'Keeffe, 66–79. Abingdon: Routledge. https://doi.org/10.4324/9780203856949.ch6

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

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

Monzó Nebot, Esther. 2008. "Corpus-based Activities in Legal Translator Training." *The Interpreter and Translator Trainer* 2 (2): 221–251. https://doi.org/10.1080/1750399X.2008.10798775

Mori, Laura (ed). 2018. *Observing Eurolects. Corpus Analysis of Linguistic Variation in EU Law, Studies in Corpus Linguistics*. Amsterdam and Philadelphia: John Benjamins.

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

Pontrandolfo, Gianluca. 2012. "Legal Corpora: An Overview." *Rivista Internazionale di Tecnica della Traduzione* 14: 121–136.

Prieto Ramos, Fernando. 2014a. "Legal Translation Studies as Interdiscipline: Scope and Evolution." *Meta: Translators' Journal* 59 (2): 260–277. https://doi.org/10.7202/1027475ar

Prieto Ramos, Fernando. 2014b. "International and Supranational Law in Translation: From Multilingual Lawmaking to Adjudication." *The Translator* 20 (3): 313–331. https://doi.org/10.1080/13556509.2014.904080

Prieto Ramos, Fernando, and Diego Guzmán. 2018. "Legal Terminology Consistency and Adequacy as Quality Indicators in Institutional Translation: A Mixed-Method Comparative Study." *Institutional Translation for International Governance: Enhancing Quality in Multilingual Legal Communication*, edited by Fernando Prieto Ramos, 81–101. London: Bloomsbury.

Simonnæs, Ingrid, and Sunniva Whittaker. 2013. "The Bergen Translation Corpus *TK-NHH* – Design and Applications." *The Many Facets of Corpus Linguistics in Bergen – In Honour of Knut Hofland* (special issue of *Bergen Language and Linguistics Studies*, vol. 3), edited by Lidun Hareide, Christer Johansson, and Michael Oakes, 93–106. Bergen: University of Bergen.

Snell-Hornby, Mary. 2006. *The Turns of Translation Studies. New Paradigms or Shifting Viewpoints?* Amsterdam: John Benjamins. https://doi.org/10.1075/btl.66

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Trklja, Aleksandar, and Karen McAuliffe. 2018. "The European Union Case Law Corpus (EUCLCORP): A Multilingual Parallel and Comparative Corpus of EU Court Judgments." *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities: CRH-2*, edited by Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, 217–226. Vienna: Gerastree Proceedings.

Vanden Bulcke, Patricia. 2013. "Dealing with Deontic Modality in a Termbase: The Case of Dutch and Spanish Legal Language." *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 12: 12–32.

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

Address for correspondence

Fernando Prieto Ramos
Centre for Legal and Institutional Translation Studies (Transius)
Faculty of Translation and Interpreting
University of Geneva
Switzerland

Fernando.Prieto@unige.ch
https://orcid.org/0000-0002-4314-2813