Building representative multi-genre corpora for legal and institutional translation research

The LETRINT approach to text categorization and stratified sampling

Fernando Prieto Ramos, Giorgina Cerutti & Diego Guzmán
University of Geneva

Exploring questions of representativeness, balance and comparability is essential to tailoring corpus design and compilation to research goals, and to ensuring the validity of research results. This is especially true when the target population of texts under examination is very large and transcends a restricted area of specialization and/or covers multiple genres, as in the case of texts translated in institutional settings. This paper describes the multilayered sequential approach to corpus building applied in a comparative study on legal translation in three of these settings. The approach is based on a full mapping and categorization of institutional texts from a legal perspective; it applies an innovative combination of stratified sampling techniques integrating quantitative and qualitative criteria adapted to the research aims. The resulting corpora, categorization matrix and selection records, together with the methodological detail provided, can be useful for building other multi-genre corpora in translation studies and further afield.

Keywords: corpus, representativeness, text categorization, stratified sampling, genre, balance, legal translation, institutional translation

1. Representativeness and research needs: LETRINT's corpus-building sequence

Any corpus conceived for linguistic or translation research whose object of inquiry transcends a restricted area of specialization inevitably relies on text classification and selection for its design and compilation. Unless the target population of texts is clearly identifiable and small enough for full integration into the

https://doi.org/10.1075/ts.00014.pri
Translation Spaces 8:1 (2019), pp. 93–116. issn 2211-3711 | e-issn 2211-372x

corpus, sampling is required. To this end, explicitly addressing questions of representativeness and balance is essential to ensure the methodological soundness of corpus-based research and, hence, the validity and relevance of its results. As noted by Biber (1988, 246), corpus representativeness "determines the kinds of research questions that can be addressed and the generalizability of the results of the research."

However, there is no consensus on how to achieve representativeness. This property cannot be measured precisely but may be best viewed as a question of degrees and, as pointed out by Leech (2007, 143–144) and Zanettin (2012, 46), an aspirational goal. As regards size, seminal studies by Biber (1990, 1993) suggest that 1,000-word and 10-text samples could adequately represent the linguistic characteristics of the target population, while Oostdijk (1991) argues that samples of 20,000 words can be sufficiently representative of a genre. A growing number of authors also agree that small corpora may be adequate to analyze specialized language (e.g. Leech 1991, 10; Bowker and Pearson 2002, 48; Koester 2010, 67). Others recommend a minimum volume of several million tokens in the case of corpora compiled for the analysis of general language (e.g. Sinclair 2004, 189; Walter 2010, 429). Indeed, quantitative adequacy depends on research purpose, including, crucially, qualitative considerations. As rightly expressed by Sandra Halverson with regard to corpus use in translation research, emphasis must be put on the "relationships between various conceptions of the object of inquiry, the theories derived to explain that object, and the data and methods used to test and refine those theories" (1998, 2).

This was one of the premises that informed corpus design in the LETRINT project on institutional legal translation.[1] This paper describes how corpora were conceived and compiled to meet the needs of the project, ensuring their relevance, balance and representativeness with regard to the object of study in every phase and layer of analysis. These goals required a tailored cyclical approach that went from more general categorizations to the examination of more specific variation criteria for the adaptation of suitable compilation techniques.

The first key consideration was, of course, the research purposes. The aims and methods of investigation define both the *scope of the target population* and the level of ambition required to be able to generalize the findings of corpus analysis. In this case, the multiple goals of the LETRINT project called for the compilation of four sets of corpora of decreasing volume and increasingly rich metadata for detailed analysis in consecutive phases (see Tables 1 and 2): LINST, LETRINT 0,

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

LETRINT 1 and LETRINT 1+. Corpora were progressively refined; each set is composed of texts selected from the previous set (e.g. LETRINT 0 is a selection from LINST). As part of this sequential approach, text processing (i.e., data cleansing, verification and annotation) and available metadata increase in relation to prior sets, whereas the size of each set decreases to become a subset of the previous one. Specifically, this means that the LETRINT 1 set, of between 7.87 million and 9.31 million tokens per language, amounts to 2.19% of the LETRINT 0 set, which, in turn, represents approximately 69% of the LINST set size (see Table 1).

The first project aim was to map the scope of legal translation in international institutional settings. This entailed a massive, all-inclusive compilation of texts (the LINST corpora set) from three institutional settings where translation is instrumental in three key processes of multilingual text production from a legal perspective (law-making, implementation monitoring and adjudication), as identified in a preliminary study (Prieto Ramos 2014). The three settings selected are: the United Nations (UN), as the main umbrella intergovernmental organization, including its International Court of Justice (ICJ); the World Trade Organization (WTO), a specialized organization dealing with a broad range of trade-related issues, including regulatory, economic and technical aspects, as well as playing a dynamic dispute settlement role through its panels and Appellate Body; and the four main institutions of the European Union (EU), the world's most important supranational legal order (the European Commission, the European Parliament, the Council of the EU and the Court of Justice of the EU [CJEU]). The other three fundamental criteria considered for corpus design were:

a. time span: in order to ensure representative thematic and discursive variation in contemporary institutional translation, three entire years were selected (2005, 2010 and 2015, the year of the project launch), which means that each corpus is made of yearly "snapshot corpora" (Claridge 2008, 243) or "transversal cuts" to the patterns under scrutiny (Prieto Ramos 2004, 32), thus also enabling comparability of regular texts (such as annual reports) between sampling periods;
b. text types: all publicly available written texts from the relevant institutional repositories were considered;
c. languages of publication: also for the sake of comparability and efficiency, English, French and Spanish were selected as the languages common to the three settings (except for Spanish at the ICJ, where only English and French are official languages).

Significant technical difficulties were encountered in this phase (see more details in Section 2) before moving to the subsequent stages of text classification and sampling. The LETRINT 0 corpora set is the result of further metadata and tech-

Table 1. Overview of the LETRINT corpora

Corpora	Research aims	Main methods applied	Size per language (millions of tokens)			Size per setting in all languages (millions of tokens)		
			EN	ES	FR	EU	UN	WTO
LINST	Mapping of institutional translation	All-inclusive compilation of publicly available texts in English, French or Spanish	641.32	517.48	555.71	1,089.34	461.24	163.92
LETRINT 0	Categorization of multilingual institutional texts from a legal perspective + quantitative analysis of translation per institutional function and genre + definition of scope of institutional legal translation	Selection of translated texts according to language and technical specifications + systematic text classification	362.86	406.48	409.44	730.99	300.16	147.63
LETRINT 1	Analysis of discourse features, translation patterns and quality indicators	Stratified sampling for compilation of sampling frames (key genres as strata from each legal category) + tailored systematic sampling for selection of units within strata	7.87	8.59	9.31	9.87	9.12	6.78
LETRINT 1+	Analysis of discourse features, translation patterns and quality indicators	Targeted downsampling for annotation of genre subcorpora	N/A					

nical verifications, and categorization of all translated texts covered by the project from a legal perspective (see Section 3). This comprehensive overview was necessary to yield key quantitative results on translation volume for both primary and secondary institutional functions (see Table 3 below), as well as for specific genres within each functional category, and ultimately to further define the scope of legal translation in the three settings. In turn, this corpora set was instrumental to generating smaller corpora, LETRINT 1 and LETRINT 1+, for a more detailed analysis of discourse features, translation patterns and quality indicators. These aims required a multi-layered, multi-stage application of stratified sampling techniques in which quantitative and qualitative criteria were adapted to the size, nature and variation of each key genre (or stratum) selected from each legal category quantified in LETRINT 0 (see Section 4). From the first aligned corpus, LETRINT 1, further parallel corpora can be built for quantitative and qualitative analysis. In the case of the LETRINT project, additional targeted sampling was conducted to compile a corpus for annotation of specific variables, LETRINT 1+. This corpus will not be described in this paper. Rather, the next sections focus on the methodological aspects of design and compilation of the other corpora following the sequence presented here. The results of corpus analysis fall beyond the scope of this article.

2. Mapping institutional text production: The LINST corpora set

The LINST set compilation not only constituted a major challenge and milestone in the initial mapping of institutional translation, but also a condition for building subsequent sets. Unlike other corpora aimed at studying one or several given textual genres,[2] as mentioned above, LINST was designed to include as many genres as possible within the settings examined. This posed several data retrieval and acquisition issues. The first one was derived from the diversity of sources of the four main EU institutions comprised in the third setting of the study, which, as opposed to the UN and the WTO, required identifying several institution-specific repositories (see full list in Table 2) apart from the main common database of EU legal texts, EUR-Lex. The most significant challenge, however, was related to data acquisition, as none of the repositories provided procedures for bulk data downloading. After various consultations with institutional contacts, a decision

[2] See, for example, Trklja and McAuliffe 2018 on the EU case law corpus (EUCLCORP) composed of EU court judgments.

Table 2. Description of the LETRINT corpora

Corpora set	Main features	Source repository(-ies)	Selection criteria	Cumulative metadata	Annotation
LINST	– Three corpora – Monolingual – Comparable – Synchronic	EU: – EUR-Lex [a] – CJEU's website [b] – European Council Document Register [c] – European Parliament Public Register of Documents [d] – Register of Commission Documents [e] UN: Official Document System (ODS) repository [f] WTO: WTO Documents Online repository [g]	1. Publication date: – 01.01.05-31.12.05 – 01.01.10-31.12.10 – 01.01.15-31.12.15 2. Document availability: at least in one of LETRINT's working languages (English, French or Spanish)	1. General metadata included in all repositories: – Document title – Document symbol – Date of publication 2. Organization/repository specific metadata, i.e.: – Type of legal act (EUR-Lex) – Job number (UN) – Document synthesis (WTO) 3. LETRINT project generated metadata: – Word count for each linguistic version – Document code (same for all linguistic versions) – Organization specific metadata (i.e., duty station in the case of UN documents)	None

Table 2. *(continued)*

Corpora set	Main features	Source repository(-ies)	Selection criteria	Cumulative metadata	Annotation
LETRINT 0	– Three corpora – Monolingual – Comparable – Synchronic	LINST corpora set	1. Source language: English (French for CJEU documents) 2. Document availability: English, French and Spanish versions, except for the ICJ	1. General and organization/repository-specific metadata from LINST corpus 2. LETRINT project generated metadata: – Word counts – Text categorization from a legal perspective	None
LETRINT 1	– One corpus – Parallel – Synchronic	LETRINT 0 corpora set	1. Editorial status: final versions only 2. Quantitative and qualitative criteria (see Section 4)	1. LETRINT 0 metadata 2. Additional LETRINT project generated metadata (e.g. document status, subject, body)	Parts-of-speech (POS) tagging
LETRINT 1+	– One corpus – Parallel – Synchronic	LETRINT 1 corpus	Further quantitative and qualitative criteria	LETRINT 0/1 metadata	– POS – Manual annotation

a. https://eur-lex.europa.eu/.

b. https://curia.europa.eu/.

c. https://www.consilium.europa.eu/en/documents-publications/public-register/.

d. http://www.europarl.europa.eu/RegistreWeb/home/welcome.htm.

e. http://ec.europa.eu/transparency/regdoc/index.cfm.

f. https://documents.un.org/prod/ods.nsf/home.xsp.

g. https://docs.wto.org/.

was made to develop *ad hoc* scripts based on web crawlers for each platform.[3] These scripts were designed to siphon off all available documents and their corresponding metadata according to the selection criteria specified in Table 2. The only publicly accessible texts deliberately excluded from the project were webpages, not only because of the impracticality of retrieving these texts diachronically from all relevant websites, but also because they did not qualify as relevant to the project needs due to the fact that their content partially overlapped with other genres such as press releases, and they were peripheral with regard to core legal genres (see Section 3).

Overall, more than 800,000 files in .pdf, .doc(x), .rtf, .htm(l) and .txt formats were downloaded.[4] Following acquisition, productivity statistics provided by institutional contacts enabled us to verify proportions and validate the accuracy of the LINST data. Once downloaded, each document was assigned a unique alphanumeric code for all its linguistic versions stating: (1) its organization (EU, UN or WTO); (2) its year of publication (2005, 2010 or 2015); and (3) a unique numeric code (e.g. "WTO_2010_14524"). For ease of reference, the code remains the same in the subsequent corpora sets. Documents for which a simple text file was not available were converted to .txt format (UTF-8), with a view to: (1) conducting automatic word counts with WordSmith Tools version 6 (Scott 2012); and (2) aligning and annotating documents in later stages of the project.

Metadata were stored in a series of .csv files, which were cleansed, processed and merged into more legible .xlsx tables. Due to the divergences and gaps detected in institutional repositories, some pieces of information had to be completed or generated subsequently. This was particularly demanding, as some details could only be retrieved by checking thousands of files manually. The most critical difficulties in completing the metadata were related to source language specification, as neither the EU nor the UN repositories offer this information. In the case of the UN, the source language, usually indicated in the documents, was extracted by two means: (1) an *ad hoc* script specifically developed to this end; and, (2) by verifying each document individually. As for the EU, except for CJEU documents originally drafted in French and unless otherwise specified, it was presumed that English was the source language, even if there is officially no "original text." Once again, institutional statistics helped to validate this presumption. In this respect, it is worth

[3] The scripts were developed by Philippe Baudrion (University of Geneva), Florian Katenbrink and Aleskander Umov, to whom we are very grateful. Technical collaboration with the Centre for Trade and Economic Integration (CTEI) of the Graduate Institute of International and Development Studies (IHEID) and insights provided by the EUR-Lex Helpdesk and other institutional contacts were instrumental during this phase of the project.

[4] Documents only available in other formats were automatically dismissed.

reminding that institutional English is overwhelmingly used as a *lingua franca* by non-native speakers in these settings. It has been described as a simplified version of English with features of its own, such as cultural detachment or 'deculturalization' (van Els 2001, 329), 'permeability' to non-idiomatic uses and reproduction of conventions specific to each institution (Prieto Ramos 2014, 318; see also e.g. Husa 2012; Felici 2015 and Mori 2018 on EU English; and Steinberg 2004 and Zhao and Cao 2013 on institutional discourse at the WTO and the UN, respectively).

Despite its raw nature as a research resource due to its volume (over 1.71 billion tokens) and limited treatment, the LINST set can be exploited for various types of descriptive analysis. In the framework of the LETRINT project, it provides a full picture of textual production, reflecting the core functions shared by all the institutions, as well as the diversity of bodies involved and themes dealt with. The data also confirm the higher volume of editorial and translation work carried out by the EU institutions, as evidenced by the size of its subcorpora, which amount to approximately 63.5% of the set. Finally, LINST reasserts the predominance of English as a working language in international fora. As opposed to the subsequent parallel corpora (see Table 2), in which French and Spanish documents are always lengthier, the size of LINST monolingual subcorpora in English are remarkably larger than in the other two languages due to the proportion of texts that were not translated.

3. Categorization of institutional texts: From LINST to LETRINT 0

As outlined above, the LETRINT 0 set was designed to categorize LINST texts originally drafted in English and translated into French and Spanish. This categorization was essential to quantify and situate genres according to the institutional functions fulfilled and reflected in text production processes in the three settings. This initially required processing the LINST set components in order to merge documents that had been split (e.g. annexes separated from the main texts), and to discard: (a) texts originally drafted in any language other than English (except in the case of the CJEU), when this information was available; (b) texts not available in any of the project's working languages; (c) text duplications; and (d) incomplete files. This was done semi-automatically by analyzing the previously processed metadata, creating Excel formulas and, where necessary, conducting additional manual verifications. As also noted in the previous section, this filtering process was very significant in the case of English texts due to the use of this language as the single working language in many instances. This was rare in the case of French and very rare in the case of Spanish. In these two languages, the most relevant screening factor, which also applies to English, was the exclusion of (a) duplications of legislation found in EUR-Lex compilations and, to a much lesser extent, (b) texts issued by certain EU institutions or bodies that were not targeted for inclusion in the project but were found in EUR-Lex (e.g. the European Central Bank or the European Court of Auditors). The application of these criteria led to the exclusion of approximately 169.5 million tokens in the EU setting (53.4 in English, 57.6 in French and 58.5 in Spanish). Files excluded for strictly technical reasons represented an insignificant proportion, but their removal was necessary to avoid data distortions in the next stages.

The full categorization of texts according to the three categories outlined in Section 1 was based on a previous legal contextualization of institutional missions and text production processes (Prieto Ramos 2014, 2017). A cyclical approach was adopted in order to situate genres with regard to the three main functional categories mentioned above. This meant that the boundaries and internal structure of these categories were gradually refined during the classification process considering new insights yielded by the corpus itself (see more details in Prieto Ramos, 2019).

Each textual unit could only be included in one category and subcategory. This entailed filtering or completing metadata (for example, no text typology information was available in the UN repository) and, where necessary, verifying titles, content and discourse features in order to identify text functions. Several validators (at least two LETRINT researchers per organization, including the project supervisor) were involved in the process. Classification issues were compounded by the interconnection between legal categories, for instance, in the case of texts that deal with implementation matters but may also be used as preparatory documents in law- or policy-making. In turn, the boundaries between soft law and (equally non-binding) policy-making texts are not always clear-cut. It was finally decided to merge them into a single subcategory of "soft law and other policy formulation" within "law- and policy-making."

While the multifunctional nature of texts is acknowledged in the project, the main legal function of each text prevailed for classification purposes, and representative genres were selected for further analysis considering this key functional dimension (see Section 4.1). Furthermore, the fact that classification criteria are explicit, traceable and systematically applied by several validators (see further details in Prieto Ramos, 2019) means that other researchers may access and adopt this classification, or eventually reorganize or further define subcategories depending on their research focus and approach. As a final methodological caveat, it is also worth noting that there was a significant risk of duplications in the case of the EU legal acts, since they are elaborated by several institutions as part of the ordinary legislative procedure, involving multiple drafts and stages before final approval. To offset this risk, all EU databases and metadata were verified with the utmost care to make sure that each draft or working document was registered only once.[5]

Table 3. LETRINT text categorization matrix (from Prieto Ramos, 2019, 40)

Main functional categories	**Subcategories based on relevance to main function (illustrative genres)**
1. Law- and policy-making	
1.1. Hard law	a. Key (e.g. treaties, agreements, regulations, directives) b. Secondary (input, instrumental or derived) (e.g. technical reports, proposals, minutes)
1.2. Soft law and other policy formulation	a. Key (e.g. declarations, resolutions, guidelines, model laws) b. Secondary (input, instrumental or derived) (e.g. records, technical reports, letters)
2. Monitoring	
2.1. Mandatory compliance monitoring	a. Key (e.g. States' reports, monitoring bodies' reports) b. Secondary (input, instrumental or derived) (e.g. procedural notes, letters)
2.2. Pre-accession monitoring	a. Key (e.g. communications, questions and replies) b. Secondary (input, instrumental or derived) (e.g. statements, minutes)
2.3. Other monitoring and implementation matters	a. Key (e.g. progress reports, working papers, notes) b. Secondary (input, instrumental or derived) (e.g. checklists, letters)
3. Adjudication	a. Key (primary case documents, e.g. requests, appeals, judgments) b. Secondary (input, instrumental or derived) (e.g. activity reports, summaries, press releases)
4. Administrative functions (not included in other categories)	
a. Organization's human resources, finance and procurement	(e.g. budgets, recruitment notices, calls for tenders, staff notices)
b. Other coordination and internal matters	(e.g. minutes, notes, presentations, reports)

[5] Apart from drafts, other information stored in the metadata includes: corrigenda, revisions and final versions. This was very relevant for the analysis of translation processes and volumes.

Insights gained through the classification work supported adjustments to the text categorization matrix, as reflected in Table 3, including an additional category of "administrative functions," which can be considered instrumental to the other main functional categories. The distinctions between key and secondary subcategories within functional categories (i.e., input or preparatory, instrumental or derived texts) were also refined, although they were ultimately deemed irrelevant in the case of housekeeping administrative texts of the last category. Overall, these subcategories form systems of interrelated genres that gravitate around the key genres within the legal hierarchy of each organization (see Figure 1).

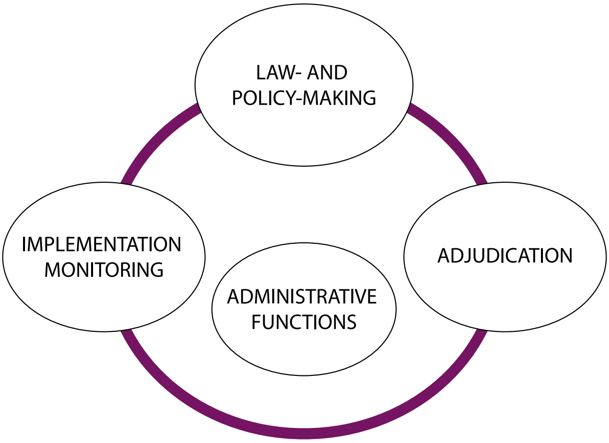


Figure 1. LETRINT primary functional categories (Prieto Ramos, 2019, 41)

Developing this text categorization matrix was a very lengthy process that, nonetheless, proved pivotal to ensuring comparability, balance and representativeness in corpus design. As noted by McEnery et al. (2006, 16), "work in text typology – classifying and characterizing text categories – is highly relevant to any attempt to achieve corpus balance" (see also e.g. Atkins et al. 1992, 14; Biber 1993, 244; McEnery and Hardie 2012, 10–11). The far-reaching data sets of LETRINT 0 provide a unique comparative picture of the components and volumes of institutional translation, and therefore the basis for investigation into the confines and features of legal institutional translation, as it appropriately frames the selection of representative genres for further empirical analysis. Furthermore, although LETRINT 0 lacks POS tagging, a script was developed to enhance its usability and index its contents. These features facilitate the re-use of this translation-oriented resource, which is particularly well suited for monolingual analyses or multilingual analyses with non-aligned corpora. The texts of LETRINT 0 can also be recycled and used to create *ad hoc* (Aston 1999) or disposable (Varantola

2000) corpora for specific research purposes, such as in the lexicometric study conducted by Prieto Ramos and Guzmán (2018) as part of the terminological strand of the LETRINT project.

4. Stratified systematic sampling: From LETRINT 0 to LETRINT 1

As pointed out by McEnery et al. (2006, 20), "[o]nce the target population and the sampling frame are defined, different sampling techniques can be applied to choose a sample which is as representative as possible of the population." Given the ambitious scope of the LETRINT corpora, stratified sampling was adopted as the most suitable technique to ensure representativeness and balance in the selection of textual units of key genres of each legal category. As noted by Biber (1993, 244), "stratified samples are almost always more representative than non-stratified ones (and they are never less representative)." This method ensures that "various levels of a population are represented, even in a small sample and even if some of the levels are minorities" (Mellinger and Hanson 2017, 12).

Indeed, sampling by strata (genres in this case) would not only avoid the risks of accidentally overlooking representative units through simple random sampling, but would also enable a tailored sampling approach by subgroup in order to further reinforce representativeness. This has been similarly recognized in statistics: "since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population" (European Commission 2014, 4). In the case of LETRINT, considering that it was impossible to include all textual units of a single genre, institution and year, systematic sampling was applied according to quantitative and qualitative criteria within the selected subgroups in stratified sampling.

4.1 Selection of genres

The first step was the selection of genres (or subcategories) as sampling frames from each main functional category (see Table 3). As mentioned above, the preliminary mapping of the first phase enabled this task. Given the project aims, the focus was put on key genres that perform the core institutional legal functions under examination (e.g. directives and regulations in law-making, implementation reports in monitoring and judgments in adjudication – see Table 4). In line with Sinclair's (2005, 4) recommendations, in order to improve balance, comparability and representativeness:

a. as a rule, two genres from each category and setting[6] were selected considering their volumes, legal relevance, comparability and complementary nature, for example, samples from the most significant body of law of each setting (binding decisions at the WTO, resolutions and ILC reports from the UN, and both binding and non-binding legal acts from the EU), national reports and international body reports within monitoring procedures, and documents initiating or closing court or adjudicative proceedings; and
b. further validation ensured that selected genres were both proportionally significant to represent a particular category and viable for sampling purposes, i.e., they include sufficient textual units in the years covered by the project.[7]

As can be observed in Table 4, in dealing with internal genre diversity, a pragmatic inclusive approach was adopted to identify genres sharing the same functional features (e.g. binding decisions or dispute settlement reports at the WTO) and genre internal subgroups according to body or textual subtype (see Section 4.2), rather than dividing genres into entirely separate subcategories on this basis.

In terms of comparability, the selection of monitoring and adjudication genres produced by Member States and institutional bodies in the three settings would enable comparative linguistic analyses per group of drafters. In the case of the ICJ, however, this required special attention to also ensure representativeness: proceedings' initial documents (applications instituting proceedings) and closing documents (judgments or advisory opinions) were not selected due to insufficient texts (applications) in the three years or because they were co-drafted in the two official languages and cannot be studied as translations. Parties' pleadings (memorials, counter-memorials, replies and rejoinders in contentious cases) and judges' opinions (separate and dissenting opinions) were selected instead as representative of legal argumentation by the parties and the Court, respectively. Finally, as regards law-making, the EU was the only context where both hard law and soft law had significant volumes to be represented in LETRINT 1, as opposed to only soft law at the UN and hard law at the WTO. This reflects the nature of the most prominent law-making activities in each setting (see Prieto Ramos 2017 for further legal contextualization) and does not compromise comparability in corpus design.

[6] Law-making at the WTO was an exception in that only one genre was selected: binding decisions of the highest decision-making bodies.

[7] In some exceptional cases, however, only one document of a particular key genre was available in one of the years analyzed (e.g. WTO Appellate Body reports in 2010), and, in one case, no documents were available (States' pleadings at the ICJ in 2005, within an otherwise key genre whose inclusion was mandatory for qualitative reasons, as noted in Prieto Ramos 2017, 190–191).

Table 4. Subcategories selected for stratified sampling in LETRINT 1

	Law- and policy-making (1.1-A & 1.2-A)[a]		Mandatory compliance monitoring (2.1-A)		Adjudication (3-A)	
EU	– Regulations – Directives – Recommendations – Guidelines – Opinions		– European Commission (EC) reports – Prior notifications of concentrations issued by the Directorate-General for Competition of the EC – EC invitations to submit comments on State aid pursuant to art. 88(2) of the EC Treaty		– CJEU judgments – Actions before the CJEU[b]	
	LETRINT 1 sample size:[c]	1.47	LETRINT 1 sample size:	0.34	LETRINT 1 sample size:	1.17
	1.1-A & 1.2-A in LETRINT 0:	33.94	2.1-A in LETRINT 0:	4.50	3-A in LETRINT 0:	20.80
UN	– Resolutions adopted by the General Assembly, the Security Council, the Economic and Social Council (ECOSOC), and the Human Rights Council (HRC) – Reports issued by the International Law Commission (ILC)		– National reports in selected mandatory monitoring procedures[d] – Concluding observations issued by the UN human rights treaty bodies – Decisions, views and opinions issued by selected UN treaty bodies[e] in complaint procedures		– Separate and dissenting opinions by ICJ judges – States' pleadings[f] in contentious cases before the ICJ	
	LETRINT 1 sample size:	0.59	LETRINT 1 sample size:	1.81	LETRINT 1 sample size:	0.46
	1.1-A & 1.2-A in LETRINT 0:	1.86	2.1-A in LETRINT 0:	17.16	3-A in LETRINT 0:	1.95
WTO	– Decisions by the Ministerial Conferences and binding decisions[g] by the General Council		– Reports issued by the Trade Policy Review Mechanism (TPRM) Secretariat – National reports presented as part of Trade Policy Reviews (TPR)		– Dispute settlement reports issued by panels and the Appellate Body – States' requests to the Dispute Settlement Body (DSB) for the establishment of a panel	
	LETRINT 1 sample size:	0.03	LETRINT 1 sample size:	1.03	LETRINT 1 sample size:	0.95
	1.1-A & 1.2-A in LETRINT 0:	0.13	2.1-A in LETRINT 0:	14.81	3-A in LETRINT 0:	7.83

a. See full set of categories in Table 3.

b. Including direct actions, references for a preliminary ruling and appeals.

c. In millions of tokens in the source language only.

d. Including procedures before the UN human rights treaty bodies, selected committees of the Security Council, and the Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapon (NPT).

e. Including the Committee against Torture (CAT) and the Human Rights Committee (CCPR).

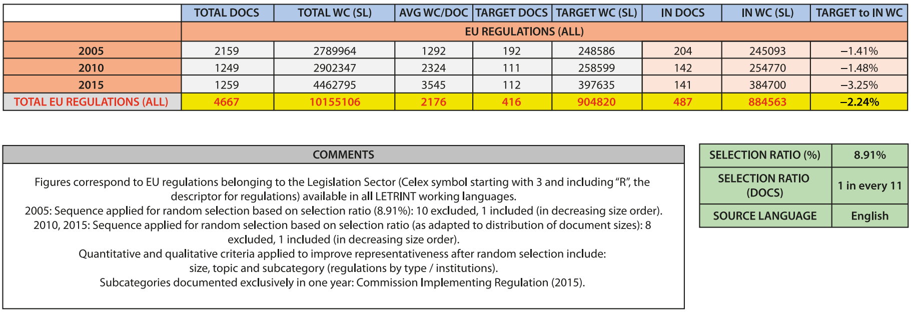
f. Including memorials, counter-memorials, replies and rejoinders.

g. Including decisions on accession to the WTO, decisions on schedules of concessions, decisions on the TRIPS Agreement, procedural decisions and waiver decisions.

4.2 Tailored selection of textual units by stratum

After the selection of genres, systematic sampling techniques were applied according to relevant quantitative and qualitative criteria in several stages. The first one consisted of setting the *sampling intervals for systematic selection* of units from each sampling frame, i.e., deciding the proportion to be selected from each genre and setting minimum and maximum thresholds in terms of tokens. In order to strike a balance between quantitative adequacy, corpus manageability and relevance to the project needs, a maximum of approximately 7 to 10 million words per institutional setting (including the three languages) was established as a target for each subcorpus, but it was decided that the internal distribution and volume of corpus components would be adapted to the significance of functional categories in each setting on the basis of the previous mapping (see Section 2).

Accordingly, the sampling target was set at one third of total text volume per genre and year, with a cumulative minimum threshold of 28,000 tokens and a maximum threshold of approximately 900,000 tokens[8] for the three years examined in the project. The standard interval for selection was therefore one out of every three texts from the sampling list of textual units. However, in the case of large volumes, the sampling interval was adjusted according to the maximum threshold of 900,000 tokens. For example, in the case of EU regulations, when this cap was applied to the total volume of almost 10.16 million tokens, the resulting selection ratio was 8.91% of texts and the target interval was one in every eleven texts (see Figure 2).



	TOTAL DOCS	TOTAL WC (SL)	AVG WC/DOC	TARGET DOCS	TARGET WC (SL)	IN DOCS	IN WC (SL)	TARGET to IN WC
	EU REGULATIONS (ALL)							
2005	2159	2789964	1292	192	248586	204	245093	−1.41%
2010	1249	2902347	2324	111	258599	142	254770	−1.48%
2015	1259	4462795	3545	112	397635	141	384700	−3.25%
TOTAL EU REGULATIONS (ALL)	4667	10155106	2176	416	904820	487	884563	−2.24%

COMMENTS

Figures correspond to EU regulations belonging to the Legislation Sector (Celex symbol starting with 3 and including "R", the descriptor for regulations) available in all LETRINT working languages.
2005: Sequence applied for random selection based on selection ratio (8.91%): 10 excluded, 1 included (in decreasing size order).
2010, 2015: Sequence applied for random selection based on selection ratio (as adapted to distribution of document sizes): 8 excluded, 1 included (in decreasing size order).
Quantitative and qualitative criteria applied to improve representativeness after random selection include:
size, topic and subcategory (regulations by type / institutions).
Subcategories documented exclusively in one year: Commission Implementing Regulation (2015).

SELECTION RATIO (%)	8.91%
SELECTION RATIO (DOCS)	1 in every 11
SOURCE LANGUAGE	English

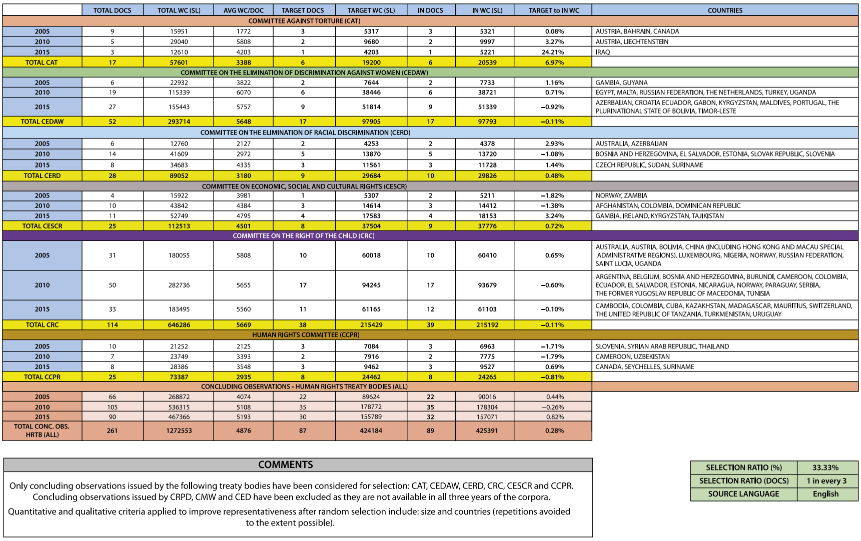
Figure 2. Selection record (overview page) for EU regulations (law-making)

As a result of this sampling process, the EU subcorpus of LETRINT 1 is the largest and the WTO subcorpus is the smallest, but the focus on key genres partly offsets the more important quantitative differences found in the previous LETRINT cor-

[8] This number of words could be exceeded slightly depending on specific document sizes.

pora (see Table 1), due to the thresholds applied to enhance comparability, balance and representativeness (see Table 4). In other words, the smaller the subcorpus, the higher the proportion of the total volume of institutional translation represented in LETRINT 1 after exclusion of less relevant genres. This balancing also illustrates that representativeness is not necessarily at odds with comparability in building a multi-institutional, multi-genre corpus, since it prevents mutually conflicting distortions (a concern raised by Leech 2007, 142).

Considering the quantitative sampling goal, as well as the number and average length of textual units of each selected genre, the target number of documents and words were automatically calculated and integrated into *selection records* by genre or subcategory (see illustrative Figures 2 and 3). The sampling interval was applied in descending size order, always excluding the first (largest) units. In the case of genre populations that include a high number of units and no major deviations from the average text length, this systematic sampling yielded results within the expected target. However, in the case of short sampling lists (i.e., few textual units in a genre and year) or significant variations between textual unit sizes (i.e., acute deviations within a sampling frame), systematic sampling required additional modulation.



	TOTAL DOCS	TOTAL WC (SL)	AVG WC/DOC	TARGET DOCS	TARGET WC (SL)	IN DOCS	IN WC (SL)	TARGET to IN WC	COUNTRIES
COMMITTEE AGAINST TORTURE (CAT)									
2005	9	15951	1772	3	5317	3	5321	0.08%	AUSTRIA, BAHRAIN, CANADA
2010	5	29040	5808	2	9680	2	9997	3.27%	AUSTRIA, LIECHTENSTEIN
2015	3	12610	4203	1	4203	1	5221	24.21%	IRAQ
TOTAL CAT	17	57601	3388	6	19200	6	20539	6.97%	
COMMITTEE ON THE ELIMINATION OF DISCRIMINATION AGAINST WOMEN (CEDAW)									
2005	6	22932	3822	2	7644	2	7733	1.16%	GAMBIA, GUYANA
2010	19	115339	6070	6	38446	6	38721	0.71%	EGYPT, MALTA, RUSSIAN FEDERATION, THE NETHERLANDS, TURKEY, UGANDA
2015	27	155443	5757	9	51814	9	51339	–0.92%	AZERBAIJAN, CROATIA ECUADOR, GABON, KYRGYZSTAN, MALDIVES, PORTUGAL, THE PLURINATIONAL STATE OF BOLIVIA, TIMOR-LESTE
TOTAL CEDAW	52	293714	5648	17	97905	17	97793	–0.11%	
COMMITTEE ON THE ELIMINATION OF RACIAL DISCRIMINATION (CERD)									
2005	6	12760	2127	2	4253	2	4378	2.93%	AUSTRALIA, AZERBAIJAN
2010	14	41609	2972	5	13870	5	13720	–1.08%	BOSNIA AND HERZEGOVINA, EL SALVADOR, ESTONIA, SLOVAK REPUBLIC, SLOVENIA
2015	8	34683	4335	3	11561	3	11728	1.44%	CZECH REPUBLIC, SUDAN, SURINAME
TOTAL CERD	28	89052	3180	9	29684	10	29826	0.48%	
COMMITTEE ON ECONOMIC, SOCIAL AND CULTURAL RIGHTS (CESCR)									
2005	4	15922	3981	1	5307	2	5211	–1.82%	NORWAY, ZAMBIA
2010	10	43842	4384	3	14614	3	14412	–1.38%	AFGHANISTAN, COLOMBIA, DOMINICAN REPUBLIC
2015	11	52749	4795	4	17583	4	18153	3.24%	GAMBIA, IRELAND, KYRGYZSTAN, TAJIKISTAN
TOTAL CESCR	25	112513	4501	8	37504	9	37776	0.72%	
COMMITTEE ON THE RIGHT OF THE CHILD (CRC)									
2005	31	180055	5808	10	60018	10	60410	0.65%	AUSTRALIA, AUSTRIA, BOLIVIA, CHINA (INCLUDING HONG KONG AND MACAU SPECIAL ADMINISTRATIVE REGIONS), LUXEMBOURG, NIGERIA, NORWAY, RUSSIAN FEDERATION, SAINT LUCIA, UGANDA
2010	50	282736	5655	17	94245	17	93679	–0.60%	ARGENTINA, BELGIUM, BOSNIA AND HERZEGOVINA, BURUNDI, CAMEROON, COLOMBIA, ECUADOR, EL SALVADOR, ESTONIA, NICARAGUA, NORWAY, PARAGUAY, SERBIA, THE FORMER YUGOSLAV REPUBLIC OF MACEDONIA, TUNISIA
2015	33	183495	5560	11	61165	12	61103	–0.10%	CAMBODIA, COLOMBIA, CUBA, KAZAKHSTAN, MADAGASCAR, MAURITIUS, SWITZERLAND, THE UNITED REPUBLIC OF TANZANIA, TURKMENISTAN, URUGUAY
TOTAL CRC	114	646286	5669	38	215429	39	215192	–0.11%	
HUMAN RIGHTS COMMITTEE (CCPR)									
2005	10	21252	2125	3	7084	3	6963	–1.71%	SLOVENIA, SYRIAN ARAB REPUBLIC, THAILAND
2010	7	23749	3393	2	7916	2	7775	–1.79%	CAMEROON, UZBEKISTAN
2015	8	28386	3548	3	9462	3	9527	0.69%	CANADA, SEYCHELLES, SURINAME
TOTAL CCPR	25	73387	2935	8	24462	8	24265	–0.81%	
CONCLUDING OBSERVATIONS - HUMAN RIGHTS TREATY BODIES (ALL)									
2005	66	268872	4074	22	89624	22	90016	0.44%	
2010	105	536315	5108	35	178772	35	178304	–0.26%	
2015	90	467366	5193	30	155789	32	157071	0.82%	
TOTAL CONC. OBS. HRTB (ALL)	261	1272553	4876	87	424184	89	425391	0.28%	

COMMENTS

Only concluding observations issued by the following treaty bodies have been considered for selection: CAT, CEDAW, CERD, CRC, CESCR and CCPR. Concluding observations issued by CRPD, CMW and CED have been excluded as they are not available in all three years of the corpora.

Quantitative and qualitative criteria applied to improve representativeness after random selection include: size and countries (repetitions avoided to the extent possible).

SELECTION RATIO (%)	33.33%
SELECTION RATIO (DOCS)	1 in every 3
SOURCE LANGUAGE	English

Figure 3. Selection record (overview page) for human rights treaty bodies' concluding observations (monitoring)

This was the main purpose in the next stage: a double process of *balancing text sizes and verification of representativeness according to additional qualitative criteria* relevant to the nature and internal variation of each genre. As regards size adjustments, the selection by sampling intervals was manually adapted to ensure representation of various sizes in the population while keeping the sample within the total quantitative target (for instance, by excluding a previously selected unit and replacing it with the next one in the sampling list when this could offset an undesirable quantitative deviation). Text fragmentation, however, was considered unnecessary to ensure representativeness and thus avoided through this process. Simultaneously, even in the case of satisfactory quantitative results in the previous sampling stage, an analysis of variation parameters was conducted to elucidate relevant qualitative criteria to enhance representativeness. These included, crucially: (a) genre subgroups according to institutional body, proceeding or textual subclassification (e.g. types of directives, treaty implementation monitoring bodies in the case of UN treaty body reports, types of proceedings within court judgments) (see internal breakdown for some genres in Table 3 and its footnotes); (b) themes; and, in monitoring and adjudication genres, also (c) countries involved (see example in Figure 3).

For the sake of sampling efficiency, these variation parameters were tailored to the nature of each genre, i.e., the approach was modulated by subpopulation to let the corpus "show the way," especially considering that the hybridity of institutional legal discourses (in particular, the use of legal and other specialized terminology) is one of the key aspects analyzed in the project. While subgroup analysis would ensure the balanced inclusion of representative procedural patterns and other legal dimensions of variation, verifications of thematic diversity and countries involved would be critical for the subsequent examination of specialized discourses and references to national legal systems. As in the case of size balance, smaller subgroups (in terms of textual units) posed the greatest challenges when applying the above auxiliary variables, and required heightened attention in order to prevent distortions. This modulation process often involved cross-checking information from institutional sources, for instance, subject metadata registered in EUR-Lex or statistics on court proceedings or subjects in official reports.

Given the differentiated language regime of the CJEU and, more specifically, the fact that French is the language of deliberations and drafting of judgments (the most prominent adjudication genre), an additional preliminary filter was applied to CJEU genres in order to align the samples as much as possible with the central language combination of the project (i.e., translation from English as the most frequent practice in the other institutions): the selection only included judgments available in all the languages and in which English is the only authentic language of the case, as well as actions and appeals in which English is the only authentic language of the case (and which, as opposed to judgments, were originally drafted in English). This approach is meant to reinforce the comparability of discourse features and translation patterns between these genres and with genres of the same category selected from the other settings.

With a view to ensuring transparency, traceability and usefulness, the criteria and results of the above tailored sampling were also included in the selection record for each genre or subcategory. As McEnery et al. (2006, 18) point out, it is important to "document corpus design criteria explicitly and make the documentation available to corpus users so that the latter may make appropriate claims on the basis of such corpora and decide whether or not a given corpus will allow them to pursue a specific research question." LETRINT's selection records are, in fact, data sets that include:

- an overview of the sampling results and criteria (see comments in the illustrative records in Figures 2 and 3);
- the full list of documents included and excluded from the sampling frame, and their metadata, including those corresponding to qualitative criteria considered in each case (subject matter, countries involved, legal procedure, etc.); and
- further sheets containing any complementary calculations, test results or descriptive statistics used to support sampling decisions.

These records can thus help other researchers to decide to what extent the LETRINT corpora meet their particular needs and how to broaden or reduce each sample where relevant.

All in all, external criteria prevailed over linguistically-defined internal criteria (see distinction in e.g. Biber 1993, 243 and McEnery et al. 2006, 14) in light of the research aims. Indeed, focusing on internal regularities and indicators such as "closure" (e.g. McEnery and Wilson 2001, 166) or "lexical density" (e.g. Corpas Pastor and Seghiri Domínguez 2007) would not have been as reliable as external variation factors, such as themes and procedures, with a view to describing the hybridity of institutional discourses and related translation issues.

Finally, on a technical note, it is worth mentioning that LETRINT 1 documents were subject to trilingual alignment and POS tagging using customized scripts based on LF Aligner[9] and TreeTagger,[10] respectively. Each of the more than 2,600 aligned documents of this corpus (a total of 7,918 texts) was individually reviewed and validated by several members of the LETRINT team to ensure the quality of the alignment and manually correct alignment problems when necessary. Unlike its predecessors, the LETRINT 1 corpus will be uploaded into a customized online concordancing tool for trilingual quantitative and qualitative

[9] https://sourceforge.net/projects/aligner/.

[10] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

analysis chosen on the basis of a previous comparison of available tools (see Cerutti 2017). The LETRINT 1 corpus will serve as a stepping stone to move into LETRINT 1+, which will be composed of a selection of LETRINT 1 texts and will include rich annotation (including discourse features and translation difficulties). Both resources will have the potential to underpin the quantitative or qualitative analysis of a wide range of features (discursive, lexical, structural, semantic, thematic, terminological, etc.) and variables (textual genre, legal functional category, period, institution, etc.) that are relevant not only for legal and institutional translation studies, but also for legal, linguistic or discourse studies (both monolingual or multilingual) in other related areas.

5. Concluding remarks

The LETRINT multilayered sequential approach to corpus design illustrates the value of considering quantitative and qualitative parameters to ensure balance, representativeness and comparability in meeting research needs, and thus challenges reductionist ideas about the adequate size of specialized corpora. This can only be assessed in conjunction with the qualitative adequacy of the corpus components and attributes. In this case, the level of ambition of the project goals is reflected in the resulting LETRINT corpora sets, of unprecedented scope and internal granularity in the field of legal and institutional translation. This collection of corpora is unique not only because it is based on a first-of-its-kind mapping and categorization of all multilingual text production in three international institutional settings over three years (resulting in the massive LINST corpora set and the LETRINT 0 set), but also because of the innovative combination of sampling techniques tailored to the subsequent aims of analysis of discourse features and translation patterns (leading to the parallel LETRINT 1 corpus and its derived LETRINT 1+ corpus).

As opposed to other multi-genre translation-driven corpora where the target populations are presented as a given from the outset, the challenge of defining the fuzzy boundaries and internal structure of legal translation within institutional translation (the central aim of LETRINT's first phase) implied that the selection of representative genres for stratified sampling had to be grounded on several interconnected empirical foundations (from more general to more specific): legal contextualization of processes of text production, categorization of texts into key and secondary genres from a legal perspective, and consideration of the translation volume, legal relevance, comparability and complementary nature of the genres elicited.

Given the aims of the subsequent phases of the project (i.e., analysis of discourse features, translation patterns and quality indicators), the other major challenge for corpus design was related to the hybrid nature of the object of study, since institutional legal translation in a broad sense, like legal translation more generally, deals with heterogeneous discourses that are shaped by multiple specialized subjects, procedures and situational variables. In order to reflect internal variation, avoid undesirable deviations and therefore reinforce representativeness, systematic sampling at regular intervals within each genre population was accordingly followed by a double process of text size adjustments and application of additional (external) qualitative criteria as auxiliary variation parameters. These criteria (including institutional bodies, proceedings, themes and countries involved) were modulated by text subpopulation, so that the corpus could yield relevant data by genre, rather than associating representativeness with internal linguistic regularities. This modulation by stratum proved one of the key advantages of stratified sampling, and also showed that comparability can be preserved while enhancing representativeness through the process.

Overall, LETRINT's multi-layered, multi-stage approach confirms that "the design of a representative corpus is not truly finalized until the corpus is completed, and analyses of the parameters of variation are required throughout the process of corpus development in order to fine-tune the representativeness of the resulting collection of texts" (Biber 1993, 256), and this could only be achieved in a "cyclical fashion" (*ibid*) through each stage of the corpus-building sequence. The corpus-based matrix designed for institutional text categorization, as well as the multi-componential selection records created to trace all sampling criteria and results, are further innovations introduced by the project. Furthermore, not only are these instruments and the LETRINT corpora reusable for other studies on legal or institutional translation, but the whole methodological approach explained here could prove beneficial to multi-genre corpus designers in other areas of translation and linguistic research.

References

Aston, Guy. 1999. "Corpus Use and Learning to Translate." *Textus* 12: 289–314.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. "Corpus Design Criteria." *Literary and Linguistic Computing* 7 (1): 1–16. https://doi.org/10.1093/llc/7.1.1

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, Douglas. 1990. "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation." *Literary and Linguistic Computing* 5: 257–269. https://doi.org/10.1093/llc/5.4.257

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

Bowker, Lynne, and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London and New York: Routledge. https://doi.org/10.4324/9780203469255

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 & Timme.

Claridge, Claudia. 2008. "Historical Corpora." *Corpus Linguistics*, edited by Anke Lüdeling, and Merja Kytö, 242–259. Berlin: Mouton de Gruyter.

Corpas Pastor, Gloria, and Miriam Seghiri Domínguez. 2007. "Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor [Establishing a corpus representativeness threshold through the N-Cor algorithm]." *Procesamiento del lenguaje natural* 39: 165–172.

European Commission. 2014. "Theme: Sample Selection–Main Module." *Memobust Handbook on Methodology of Modern Business Statistics*. Brussels: European Commission. Accessed December 18, 2018. https://ec.europa.eu/eurostat/cros/system/files/Sample%20Selection-01-T-Main%20Module%20v1.0_1.pdf

Felici, Annarita. 2015. "Translating EU Legislation from a 'Lingua Franca': Advantages and Disadvantages." *Language and Culture in EU Law: Multidisciplinary Perspectives*, edited by Susan Šarčević, 123–140. Farnham: Ashgate.

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: Translators' Journal* 43 (4): 494–514. https://doi.org/10.7202/003000ar

Husa, Jaakko. 2012. "Understanding Legal Languages-Linguistic Concerns of the Comparative Lawyer." *The role of legal translation in legal harmonization*, edited by Cornelis J. W. Baaij, 161–181. The Hague: Kluwer Law International.

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

Leech, Geoffrey. 1991. "The State of the Art in Corpus Linguistics." *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, edited by Karin Aijmer, and Bengt Altenberg, 8–29. London: Longman.

Leech, Geoffrey. 2007. "New Resources, or Just Better Old Ones? The Holy Grail of Representativeness." *Corpus Linguistics and the Web*, edited by Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, 133–149. Amsterdam: Rodopi. https://doi.org/10.1163/9789401203791_009

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge and New York: Cambridge University Press.

McEnery, Tony, and Anita Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

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

Mellinger, Christopher D., and Thomas A. Hanson. 2017. *Quantitative Research Methods in Translation and Interpreting Studies*. London and New York: Routledge.

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

Oostdijk, Nelleke. 1991. *Corpus Linguistics and the Automatic Analysis of English*. Amsterdam and Atlanta: Rodopi.

Prieto Ramos, Fernando. 2004. *Media and Migrants: A Critical Analysis of Spanish and Irish Discourses on Immigration*. Oxford, Bern and New York: Peter Lang.

Prieto Ramos, Fernando. 2014. "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. 2017. "Global Law as Translated Text: Mapping Institutional Legal Translation." *Tilburg Law Review* 22 (1–2): 185–214. https://doi.org/10.1163/22112596-02201009

Prieto Ramos, Fernando. 2019. "Implications of Text Categorisation for Corpus-based Legal Translation Research: The Case of International Institutional Settings." *Research Methods in Legal Translation and Interpreting: Crossing Methodological Boundaries*, edited by Łucja Biel, Jan Engberg, Rosario Martín Ruano, and Vilelmini Sosoni, 29–47. London and New York: Routledge.

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.

Scott, Michael. 2012. *WordSmith Tools. Version 6*. Stroud: Lexical Analysis Software.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, John. 2005. "Corpus and Text–Basic Principles." *Developing Linguistic Corpora: A Guide to Good Practice*, edited by Martin Wynne, 1–6. Oxford: Oxbow Books.

Steinberg, Richard H. 2004. "Judicial Lawmaking at the WTO: Discursive, Constitutional, and Political Constraints." *American Journal of International Law* 98: 247–275. https://doi.org/10.2307/3176728

Trklja, Aleksandar, and Karen McAuliffe. 2018. "The European Union Case Law Corpus (EUCLCORP): A Multilingual Parallel and Comparative Corpus of EU Court Judgments (March 5, 2018)." *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.

van Els, Theo. 2001. "The European Union, its Institutions and its Languages: Some Language Political Observations." *Current Issues in Language Planning* 2 (4): 311–360. https://doi.org/10.1080/14664200108668030

Varantola, Krista. 2000. "Translators, Dictionaries and Text Corpora." *I corpora nella didattica della traduzione*, edited by Silvia Bernardini, and Federico Zanettin, 117–133. Bologna: CLUEB.

Walter, Elizabeth. 2010. "Using Corpora to Write Dictionaries." *The Routledge Handbook of Corpus Linguistics*, edited by Michael McCarthy, and Anne O'Keeffe, 428–443. Abingdon: Routledge. https://doi.org/10.4324/9780203856949.ch31

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

Zhao, Xingmin, and Deborah Cao. 2013. “Legal Translation at the United Nations.” *Legal Translation in Context: Professional Issues and Prospects*, edited by Anabel Borja Albi, and Fernando Prieto Ramos, 203–220. Frankfurt am Main: Peter Lang.

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

Co-author information

Giorgina Cerutti
Centre for Legal and Institutional Translation Studies (Transius)
Faculty of Translation and Interpreting
University of Geneva
Giorgina.Cerutti@unige.ch
https://orcid.org/0000-0002-3260-0866

Diego Guzmán
Centre for Legal and Institutional Translation Studies (Transius)
Faculty of Translation and Interpreting
University of Geneva
Diego.Guzman@unige.ch
https://orcid.org/0000-0001-8873-7119