CHEU-lex: a parallel multilingual corpus of Swiss and EU legislation

Felici, Annarita

How to cite

FELICI, Annarita. CHEU-lex: a parallel multilingual corpus of Swiss and EU legislation. In: Applied corpus linguistics, 2025, vol. 5, n° 3, p. 7. doi: 10.1016/j.acorp.2025.100151

This publication URL: https://archive-ouverte.unige.ch/unige:188069
Publication DOI: 10.1016/j.acorp.2025.100151

Applied Corpus Linguistics 5 (2025) 100151

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Articles

CHEU-lex: a parallel multilingual corpus of Swiss and EU legislation

Annarita Felici *

University of Geneva, Switzerland

ARTICLE INFO

Keywords:
Swiss legislation
EU bilateral agreements
Alignment
Annotation
Parallel corpora

ABSTRACT

This paper describes the design and construction of CHEU-lex, a parallel and comparable corpus of Swiss and European Union (EU) legislation. Data are available in the three languages of the Swiss Confederation (French, German and Italian) and include bilateral agreements between Switzerland and the EU and their reception in Swiss law. The corpus is a richly annotated multilingual resource and allows the analysis of legal language at several levels (macro-textual, lexical, morphosyntactic) and according to different perspectives (monolingual, cross-lingual, cross-textual, diachronic). The goal is to highlight key properties of CHEU-lex, discuss issues of legal corpus compilation and, finally, outline some applications for translation and legal linguistic research.

1. Introduction

CHEU-lex is a parallel and comparable corpus of Swiss legislation and European Union (EU) bilateral agreements published in the three official languages of the Swiss Confederation (French, German and Italian). Although Switzerland is not part of the EU, it is one of its major trading partners due to its geographical and cultural proximity. Political and economic relations are governed above all by bilateral instruments and since 1972 Switzerland and the EU have concluded around 20 main agreements and some 100 secondary agreements in sector-specific areas.

CHEU-lex was built on this background and on the basis of similar studies focusing on the impact of EU legislation in member countries (e.g., Biel, 2014; Mori, 2018). The corpus has two main objectives. On the one hand, it was built to study the impact of EU law on Swiss official languages and to explore language contact and the dynamics of translation in multilingual institutions. Recent studies in sociolinguistics have identified the presence of a macro-variety of legal language —namely Eurolect — generated during the multilingual drafting process and when implementing EU directives into national legislation (Mori, 2018). In the case of Switzerland, the multilingual translation process has a double impact on the final legislative text, because Swiss legislation is often translated from German into French and Italian.[1]

On the other hand, CHEU-lex aims at providing a richly annotated multilingual resource to explore legislative language at several levels (macro-textual, lexical, morphosyntactic) and according to different perspectives (monolingual, cross-lingual, cross-textual, diachronic). To the best of our knowledge, there are few multilingual legal corpora that combine broad language coverage with rich annotation scheme. The well-known Europarl (Koehn, 2005) and the JRC- Acquis (Steinberg et al., 2006) corpora were developed with computational goals in mind, such as machine translation and other NLP tasks. Both consist of parallel sentence-aligned legislative texts across many EU languages, with standardized formatting and tokenization—features that make them especially suitable for training and evaluating statistical and neural language models. More recently, the LETRINT[2] corpus was developed as a trilingual (English, French and Spanish) parallel and comparable corpus with a strong focus on text-genre. Its design, emphasizing genre diversity and different institutional settings, makes it particularly valuable for the analysis of discourse features and translation quality indicators.

As regards Swiss legislation, it is worth mentioning three corpora like our own, in the same languages, but with different scopes and annotation schemes. The Swiss Legislation Corpus (Höfler and Piotrowski, 2011) comprises contemporary statutory law of the Confederation and is annotated with both legal metadata and parts-of-speech (POS). These features support fine-grained stylistic analysis and enable systematic extraction of definitions based on recurring patterns across aligned legal provisions. The SwissAdmin corpus (Scherrer et. al., 2014) is a multilingual collection of press releases from the Swiss Federal Administration in English, French, German and Italian, specifically designed for

* Corresponding author. Faculty of Translation and Interpreting, University of Geneva, 40 Boulevard du Pont-D'Arve, Geneva 4, 1211, Switzerland.
E-mail address: annarita.felici@unige.ch.

[1] This is true above all for the Italian version, which is always a translation. The French one may be sometimes drafted or edited in parallel with the German version.

[2] LETRINT is a trilingual data set in English, French and Spanish including legal texts from the EU, UN and WTO.

https://doi.org/10.1016/j.acorp.2025.100151
Received 28 August 2024; Received in revised form 1 September 2025; Accepted 2 September 2025
Available online 4 September 2025

NLP research and parser training. It is sentence-aligned for each language pair and contains annotation of POS, grammatical function, verb valencies[3] and collocations to facilitate the training of syntactic parsers and multilingual lexicons. Finally, the Feuille federal /Bundesblatt/Foglio federale (Elminger, 2015) contains all the official publications of the Federal Gazette until 2014. Although it lacks parallel alignment, it is enriched with some linguistic annotation in German and French making it a useful resource for monolingual legal-linguistic studies such as terminological extraction or morphosyntactic variation across time.

CHEU-lex was conceived primarily to study translated language and the impact of EU legislation on Swiss law. It is made publicly available through NoSketchEngine[4] and involves several layers of annotation like a) contextual information on topic and date of publication, b) structural features of legal texts (title, preamble, articles, annexes), c) POS tags, d) sentence alignment and e) syntactic dependencies.

NoSketchEngine is an open-source version of the corpus management and corpus query software SketchEngine (Kilgarriff et al., 2014) albeit with certain functionality limitations.[5] Unlike the commercial version, which operates as a web-based service, NoSketchEngine must be downloaded from the main site, installed and hosted on a local computer or server, thereby requiring some technical expertise to set up the system and prepare the corpus.

The present paper provides details about the different stages of corpus construction while discussing issues that arose during corpus making. Drawing on multilingual, domain-specific data from the field of law, the corpus serves as a resource for translation equivalents, terminology management, legal drafting, cross-linguistic discourse analysis, and computational applications. In doing so, it demonstrates how the creation of a multilingual legal corpus—in this case, a parallel and comparable corpus of translated legal texts—provides real evidence of how language functions across disciplines in applied linguistics. Following an overview of the corpus composition and design with particular attention to how representativeness was achieved in the collection of data (Section 2), the focus turns to text pre-processing, alignment, and annotation (Section 3). Subsequently, the paper examines aspects of corpus queries and applications (Section 4) before concluding with remarks on corpus limits and availability (Section 5).

2. Data description and corpus design

The corpus comprises bilateral agreements between Switzerland and the EU in French, German and Italian, and Swiss domestic legislation implementing EU law in these same languages. Texts are organised into three corpora (one for each language), which in turn consist of two subcorpora: 1) bilateral agreements entered between Switzerland and the EU from the first agreement in 1972 to 2017, and 2) Federal legislation representing the reception of these agreements. Since the reception of the agreements into Federal legislation does not happen simultaneously, we collected only those agreements that had been implemented by the time of texts' download (January 2020). This means that at the initial stage of corpus compilation, only agreements signed up to December 2017 could be included. Exceptions were made in five cases, where relevant domestic legislation predated the first official agreement of 1972. In these instances, the Swiss Confederation had in some ways already addressed, or partially addressed, the policy later formalized in the agreements. As regards text genres, the subcorpus of bilateral agreements also includes accompanying texts such as additional protocols and letters of exchange between the parties. The domestic legislation corpus, by contrast, entails only laws and ordinances. While federal laws are parliamentary acts with general scope, ordinances are secondary legislation, primarily enacted by the Federal Council and are subordinate to these laws.

Table 1
The CHEU-lex corpus: Number of documents per language.

Language	Text genre	Texts	Tokens (as defined by SketchEngine)
German	Bilateral agreements	148	726,773
	Federal legislation	116	792,639
French	Bilateral agreements	148	903,247
	Federal legislation	116	1081,794
Italian	Bilateral agreements	148	822,414
	Federal legislation	116	939,847
	TOT	792	**5266,714**

The whole corpus consists of 444 bilateral agreements and 348 national legal acts (laws and ordinances), totalling 5, 266,714 tokens, as shown in Table 1.

From the perspective of sampling, the corpus was constructed by collecting the entire set of bilateral agreements available between Switzerland and the EU, together with their corresponding domestic legislative texts. This exhaustive, rather than selective, sampling approach enhances representativeness by capturing the full spectrum of legal instruments relevant to the bilateral relationship — including agreements, additional protocols, and letters of exchange on the one hand, and federal laws and ordinances on the other. Such comprehensive sampling reflects the methodological recommendation by Biber (1993) that representativeness in corpus design is achieved by including "texts that are typical of the target domain and that exhibit a range of variation within that domain" (p. 244). Furthermore, the decision to include multiple genres within the legislative domain provides insight into the linguistic and institutional interface between national and supranational law.

While balance in general-purpose corpora often aims for proportional representation across genres, registers, or domains (McEnery and Hardie, 2012, pp. 9–10), in this specialised legal corpus, balance is inherently shaped by the policy areas addressed in the agreements and the corresponding legal acts. As such, balance was not imposed but emerges from the institutional distribution of legislative activity across time and domains. Sinclair (2005) notes that the notion of balance in specialised corpora should be approached with caution, as balance may be domain-specific and context-sensitive (pp. 7–8). In the CHEU-lex corpus, the variability of legal topics — ranging from agriculture and transport to free movement of people, goods and service as well as mutual recognition—is traceable through macro- and micro-topic metadata derived from Fedlex,[6] ie. the official publication platform for Swiss Federal law. Table 2 lists the macro-topics as defined in the Classified Compilation of Federal legislation (SR) for domestic and international law.[7]

Corpus size was determined not by a target word count but through a corpus-driven approach responsive to the availability and relevance of authentic legal texts. As noted by Xiao (2010, p.6), corpus size and representativeness are "dialectically related" and while size may affect analytical scope, it is the relevance and internal coherence of a corpus that ultimately determines its utility in specialised studies. Similarly, Sinclair (2005) argues that "there is no simple numerical criterion for corpus size" that guarantees representativeness; rather, representativeness depends on the corpus's capacity to support generalisable, valid inferences about the domain under investigation (p. 7). In the context of

[3] Verb valency refers to the number and type of dependent arguments that the verb can have.

[4] www.sketchengine.eu/nosketch-engine/

[5] NoSketchEngine does not contain any corpora and only allows users to perform keyword extraction, wordlists, concordances and CQL queries.

[6] https://www.fedlex.admin.ch/en/home?news_period=last_day&news_pageNb=1&news_order=desc&news_itemsPerPage=10 (accessed on 17.02.2023)

[7] Texts are identified by the abbreviation SR, which stands for the French *Recueil systématique du droit federal*.

Table 2
Macro-topics as listed in Fedlex.

Federal law	International law
1.State – People –Authorities	01. International law in general
2. Private law–Administration of civil justice – Enforcement	02. Private law – Administration of civil justice– Enforcement
3. Criminal law - Administration of criminal Execution of sentences	03. Criminal law – Legal assistance
4. Education – Science – Culture	04. Education – Science – Culture
5. National defense	05. War and neutrality
6. Finance	06. Finance
7. Public works – Energy – Transport	07. Public works – Energy – Transport
8. Health – Employment – Social Security	08. Health – Employment – Social Security
9. Economy –Technical cooperation	09. Economy –Technical cooperation

specialised legal corpora, even relatively modest sizes can yield valuable insights, provided the corpus is systematically constructed and functionally relevant (Bowker and Pearson, 2002, p. 48).

Therefore, in this study, representativeness is conceptualised as a relative and context-dependent property. It is not defined in absolute statistical terms, but rather through the corpus's ability to support valid, replicable analyses of multilingual legal discourse within the Swiss–EU framework. This approach aligns with Biber, Conrad and Reppen's (1998) assertion that corpus representativeness "determines the research questions that can be addressed and the generalisability of the results" (p. 246). The inclusion of detailed metadata —genre, temporal coverage, language, text sections and legal topics—further enhances the corpus's empirical robustness and analytical flexibility, fulfilling the standards for effective corpus design as articulated in the literature (Bowker and Pearson, 2002; McEnery and Hardie, 2012; Sinclair, 2005; Xiao, 2010).

3. Corpus construction

The construction of a multilingual legal corpus presented both structural and linguistic challenges that required the integration of domain-specific tools, manual intervention, and corpus design decisions in terms of legal text typology. Three key aspects emerged as particularly relevant for the corpus and broader corpus-building practices: first, the importance of genre-aware pre-processing and metadata integration in legal documents; second, the need for manual refinement in multilingual alignment due to inherent translational divergences; and third, the limitations of standard NLP tools, which necessitated targeted adaptations to support accurate tagging and syntactic annotation in a legal context. Corpus construction comprises three stages: text pre-processing and metadata annotation (3.1), segmentation and alignment (3.2), POS tagging and syntactic annotation (3.3).

3.1. Text pre-processing and metadata annotation

Texts were downloaded from Fedlex and automatically annotated via Perl script, which extracted HTML metadata and inserted it into XML header. In addition to default structural metadata, we introduced tailored fields — 'text_id', 'text_type', 'decade_entry', 'date_entry', 'date_signature', 'date_status', 'original_text', 'topic_macro', 'topic_micro', 'type' and 'url' — to support corpus querying and classification. These were designed to reflect corpus-specific parameters while maintaining legal relevance. Structural segmentation preserved the hierarchical organization of legislative texts (e.g. ⟨title⟩, ⟨preamble⟩, ⟨body⟩), and XML tags were created to identify specific subsections such as ⟨article_title⟩ and ⟨annex_text⟩ (Table 1 in the Appendix).

The two main text types (Laws and Agreement) required a different treatment due to the complexity of headings and nested lists in Laws. Titles and subtitles were typically grouped as single segments to preserve coherence (e.g. in ⟨article_title⟩ and ⟨annex_title⟩ tags). Additionally, tables (common in annexes) were retained as plain text, while images (e.g. in institutional logos) were marked with placeholder ([IMG]) and footnotes were excluded, given their limited linguistic relevance.

Ensuring structural and tag consistency across the three languages involved both automated XML validation and targeted manual correction, especially for elements affecting segmentation and alignment. One key insight is that legal corpora require domain-specific pre-processing rules: standard web scraping and automatic annotation pipelines fail to handle nested structures, legal list formats, and document-level metadata, often resulting in large-scale inconsistencies.

3.2. Segmentation and alignment

Automatic sentence segmentation was performed using the *InterText Editor* (Vondřička, 2014), based on sentence-final punctuation. However, legislative texts include frequent abbreviations e.g. 'art.', 'bis.', 'ch.',[8] necessitating exception rules via Perl-Compatible Regular Expressions (PCRE). These rules were tailored to each language, with German requiring greater adaptation due to capitalization patterns and morphological complexity.[9]

Further rules were developed for all languages to prevent segmentation within alphabetized and/or numbered lists, and to ensure accurate splitting when list elements followed non-punctuated introductions.

Automatic sentence alignment was first performed using a chained scheme de-fr, fr-it, it-de, then corrected manually based on the *SketchEngine guidelines for m:n alignment*[10] to detect potential discrepancy between language versions. Fig. 1 shows a 1:0 alignment because the last German segment is included in the first segment of the French version. In this case the m:n alignment allowed to group multiple segments together, by converting it into 3:2 alignment. This proved essential in capturing both syntactic and semantic equivalence across languages as well as different translation styles, especially between German and the two romance language.

A major challenge in corpus alignment lies in the absence of a strict 1:1 correspondence between legal texts across languages. This limitation underscores the inadequacy of relying exclusively on automatic tools for aligning multilingual legal corpora, given the presence of structural asymmetries, translational variability, lengthy lists, untranslated items and terminology organized alphabetically.[11] Consequently, there is a clear need for hybrid approaches that integrate rule-based methods with expert human revision to ensure accuracy and reliability.

3.3. POS annotation and syntactic dependencies

POS tagging and lemmatisation were carried out on *SketchEngine* for Italian and German using respectively the Italian TreeTagger (Schmid et al., 2007), German RFTagger (Schmid and Laws, 2008) and TreeTagger v3.2[12] for French with additional preprocessing.[13]

The final POS-tagged and lemmatised corpus is a vertical or word-per-line (WPL) .vert file, in the format required by *NoSketchEngine*,

[8] E.g. Art. (Artikel/article/articolo), Art.13 bis, ch. (chapitre), which are typical of legal texts and usually end with a full stop.

[9] German required extra rules also for the treatment of nouns. As they are written with capital letters, InterText used to consider them as the beginning of a new sentence.

[10] https://www.sketchengine.eu/guide/setting-up-parallel-corpora/#tab-id-4

[11] Terminology listed in alphabetical order proved problematic during alignment, because the order of items changed from one language to the other. In the case of short lists, we grouped up the mismatched segments in a single one by creating many 1:n or m:n alignments. With longer lists, we aligned items line by line and we opted for 1:0 and 1:1 alignment to avoid very long segments.

[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[13] The French subcorpus was also initially annotated in SketchEngine with the integrated French FreeLing v.1.6 tagset, but the layout resulted very different from the other two subcorpora because of problems with clitics.

8 In Zusammenarbeit mit den Kantonen ordnet der Bundesrat schweizweit Betriebsvergleiche zwischen Spitälern an, insbesondere zu Kosten und medizinischer Ergebnisqualität.	8 En collaboration avec les cantons, le Conseil fédéral fait procéder à l'échelle nationale à des comparaisons entre hôpitaux – qu'il publie par la suite – en ce qui concerne notamment les coûts et la qualité des résultats médicaux.
Die Spitäler und die Kantone müssen dafür die nötigen Unterlagen liefern.	Les hôpitaux et les cantons doivent livrer les documents requis à cette fin.
Der Bundesrat veröffentlicht die Betriebsvergleiche.	

Fig. 1. Example of 3:2 alignment.

where each token is accompanied by a POS tag, and a lempos (consisting of the lemma, a hyphen and a one-letter abbreviation for the POS), as shown in Table 2 in the Appendix.[14] However, to address limitations of standard taggers with legislative texts, additional tagsets were introduced. A dedicated `LI` tag (i.e. list element) was used to annotate list markers, which are common in legislative texts and appear in various alphanumeric formats (e.g.: 1), a), iii), bb), 1., a., 1bis., 1a., A., Abis.; 2.12.3.1., 1/3, 4.11).

Further adaptations included an `OMISSIS` tag to mark the missing parts of the text, an `ABR` tag for abbreviations in Italian and German, and an `FW` tag for foreign words in the French tagset.

Despite language-specific customizations, the raw output required manual post-processing, as errors were both common to the three sub-corpora and to each individual language. The alphanumeric list tag (`LI`) proved problematic for all the cases where the letter stood for abbreviations (e.g. N. for 'number', 'p.' for "page" in French and Italian or "S." for *Seite,* equally meaning "page" in German). Acronyms such as "EU", "EIONET", "EWC") were often misclassified and had to be manually re-tagged as 'proper noun', according to the three different languages. By default, *NoSketchEngine* treats the corpus as case-sensitive, distinguishing between tokens like "Commission" and "commission"— a useful feature in legal texts, where capitalization often signals specific legal meanings. However, stand-alone uppercase letters (e.g. A', 'B', X', 'G') common in tables, Roman numerals and punctuation were frequently tagged inappropriately. Lowercase stand-alone letters were also misclassified: in French and German were typically tagged as symbols (SYM), while in Italian they were often misannotated as 'proper noun' (NPR) and as 'non-linguistic element' (NOCAT) .[15] These issues underscored the limited portability of standard NLP tools to legal corpora, without tailored tagsets and post-processing. Even advanced taggers require adaptation, as features typical of legislative texts—such as abbreviations, nested structures, and case-dependent meanings—often fall outside typical training data.

Although not integrated in NoSketchEngine, the three sub-corpora were additionally annotated with syntactic dependencies using the *SpaCy*[16] . The output is another .vert file that include the eight categories listed in the SpaCy documentation[17] (text, lemma, POS, tag, DEP, HEAD TEXT; HEAD POS, CHILDREN). Unlike the other annotation stages, dependency parsing was not manually corrected, resulting in potential classification mistakes, because the model was originally trained on news[18] and not on legal texts.

4. Corpus applications

The corpus is freely available online[19] and features a rich annotation scheme that support several types of linguistic and translation analysis. Its design enables explorations from three main perspectives: monolingual (bilateral agreements and/or Swiss legislation in one language), cross-lingual (bilateral agreements and/or Swiss legislation in the three languages) and cross-textual (comparative analysis of bilateral agreements and Swiss legislation in the three languages). In addition, the granularity of its metadata allows refined queries according to topic (macro and micro as defined by Fedlex), text section (title, preamble, body, annexes), text subsection (annex text and title, article text and title) date (signature, entry date, decades), source text. This flexibility supports a wide array of explorations, such as identifying translation equivalents, examining legal terminology, and studying cross-linguistic variation across languages, genres and time. To illustrate this, four key examples are presented:

(1) a CQL-based query using POS and metadata to explore modal structures (Fig. 2),
(2) parallel concordances for translation analysis (Fig. 3),
(3) a diachronic and genre-based analysis of complex prepositions and their frequency distributions (Fig. 4), and
(4) a syntactic dependency parse generated via SpaCy to examine intra-sentential structures (Table 3).

These examples illustrate the range of analysis enabled by the corpus's design, while also underscoring the need for specialized approaches in the study of legal corpora. They demonstrate how the integration of corpus linguistic tools with domain-specific legal knowledge facilitates the identification of linguistic patterns relevant to legal translation, terminology extraction, and stylistic precision.

Fig. 2 showcases a query, with CQL regular expressions, of the German periphrastic structure *sein zu* + Infinitive, a construction often used in legal text to express deontic modality. The POS annotation clearly helps to refine the queries, thus allowing to explore several types of linguistic patterns based on different criteria. The query interface's metadata drop-down menus allow researchers to constrain searches by specific decades, genre, topics, or sections of text (e.g., preambles vs. annexes) illustrating how the corpus architecture enables sophisticated, multi-parameter exploration of legal discourse.

Fig. 3 builds on this example by retrieving the corresponding structures in Italian and French, while the green arrow on the left indicates metadata about time period, the text type (agreement or laws) and section. These results single out both regular correspondences and asymmetries between source and translated legal texts, making the corpus a valuable resource for translation training and comparative linguistics.

Despite the limitations of NoSketch Engine compared to its commercial software SketchEngine, the web interface allows researchers to extract Wordlist, KeyWord list, collocations, simple and parallel

[14] By contrast, French Freeling v1.6 available for SketchEngine generated a six column .vert file, namely *token, POS tag, lempos, lemma, POS tag* and *lemma.*

[15] In Italian, letters like 'I' and 'A'—which may represent articles or prepositions—also required manual correction. Similar tagging errors occurred in French, where capital 'A' (used for the preposition *à*) was often misidentified as the verb *avoir*.

[16] https://spacy.io/

[17] https://spacy.io/usage/linguistic-features

[18] it_core_news_sm (based on the Italian *UD*), fr_core_news_md (based on the French *UD*) and de_core_news_md (based on the *TIGER* corpus). Unlike Italian and French, German does not use UPOS tags, because it was trained on the *TIGER* corpus, which has its own specific tagset. https://spacy.io/models

[19] https://transius.unige.ch/en/research/cheu-lex/home

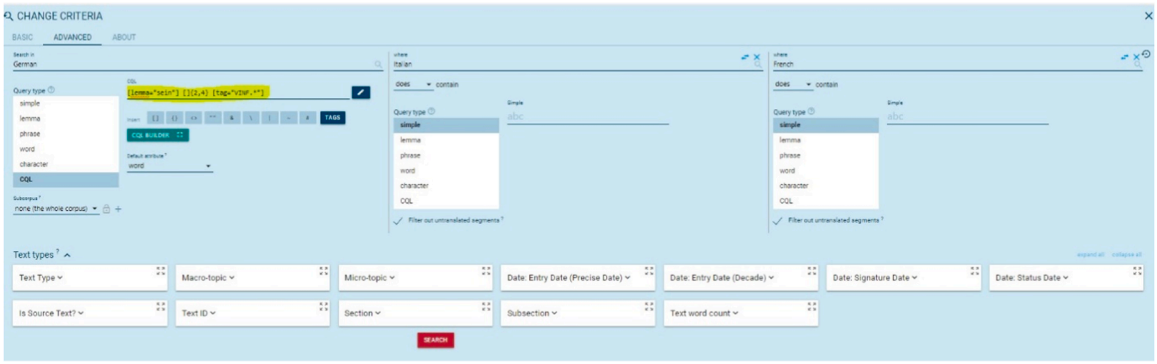

Fig. 2. Screenshot of the query [lemma="sein"][]{2,4}[tag="VINF.*"] searching for the instances of the German periphrastic pattern "sein..zu + Infinitive" in the parallel concordance.

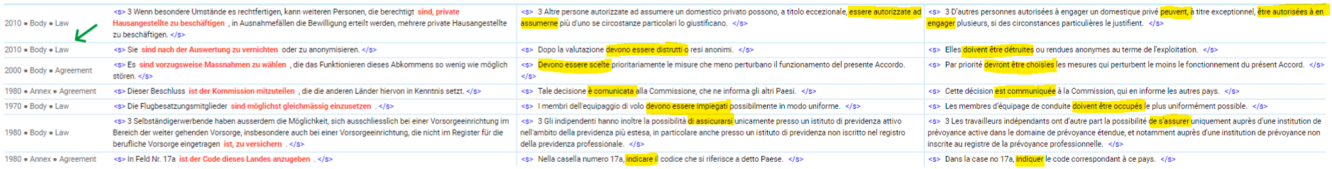

Fig. 3. Screenshot of the NoSketch Engine parallel concordance.

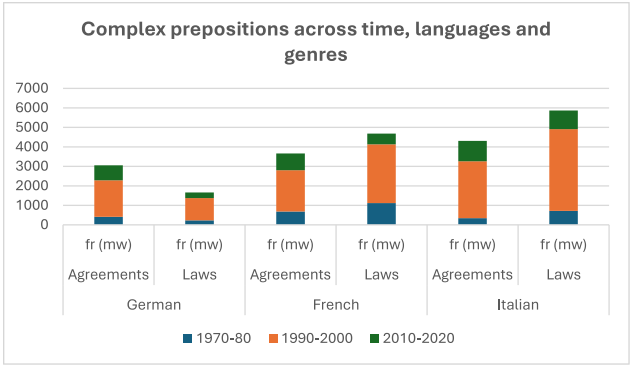

Fig. 4. Distribution of complex prepositions in the Corpus.

concordances and to display, with filtering options, more targeted results, and basic graphics. Fig. 4 shows the distribution of complex prepositions across time, languages, and genres. Complex prepositions (e.g. *Im Sinne von, en vertu de, ai sensi di*) are multi-words functional patterns that are prominent in legal texts for their cohesive function and for being influenced by legal drafting. The graph reveals for all languages and genres a peak in usage during the 1990–2000 decade, aligning with intensified legislative cooperation between Switzerland and the EU. However, if we look at their distribution across languages and genres, data suggest a higher frequency in translated language compared to the source one, thus confirming the normalization hypothesis (Baker 1996). For example, German legislative texts—usually the source—contain fewer complex prepositions than translated German agreements originally drafted in French or English. This contrast exemplifies how the corpus can be used not only to observe syntactic forms but to test hypotheses about translation universals and legal drafting conventions.

As mentioned in 3.3, the corpus was also annotated with syntactic dependencies, thus revealing syntactic relations between words that POS tagging alone cannot capture. The word that is subordinated on another is called the dependent or CHILD, and the word it depends on is the CHILD. For instance, in Table 3 "Accord" (Agreement) is identified as the sentence ROOT, i.e. the central noun around which the sentence is built. It has a modifier, "Confédération" (Confederation), which is connected by the dependency type nmod (nominal modifier), indicating that "Confédération" modifies "Accord". "Confédération" has on its turn several CHILDREN or dependents ("entre", "la", "suisse") and it is part of a larger phrase linked by the conjunction "et" (and) to "Communauté", forming a compound noun phrase with "économique européenne". Each entry in the CHILDREN column lists words that are syntactically dependent on the word in the TEXT column, showing how each word's role is interconnected to build phrases and clauses in the sentence.

Comparing the syntactic output in different languages highlights preferred linguistic choices in legal sentences, as well as potential shifts and nuance in legal meaning. Since the order of words and sentences impact on meaning, their dependencies also unveil the use of plain language and clear communication. This level of syntactic detail is essential for understanding phrase structure, coordination, and subordination in legal clauses—particularly in multi-layered noun phrases, which are very common in legislative texts. This example demonstrates the type of information *SpaCy* provides that *NoSketchEngine* does not: namely, the hierarchical structure and syntactic roles of words within sentences or valuable insight for studies on sentence clarity. As it is not integrated into NoSketchEngine and was not manually corrected, this layer is available upon request for researchers interested in syntactic structure or plain language assessment.

Together, these examples demonstrate how CHEU-lex goes beyond standard corpus functionality by offering targeted tools for linguistic, translational, and comparative legal research. Its combination of aligned multilingual texts, detailed metadata, POS tagging, and syntactic parsing supports both fine-grained and macro-level analyses. While its size (~5 million tokens) is modest by NLP standards, its rich annotation scheme and focus on parallel legal texts make it a valuable resource for applied linguistics, legal translation studies, and natural language processing. Applications range from translation pedagogy and terminology extraction to testing cross-lingual embeddings and evaluating NLP models on domain-specific language.

Table 3
Example of a Dependency parsing in the French subcorpus.

TEXT	LEMMA	POS	TAG	DEP	HEAD TEXT	HEAD POS	CHILDREN
Accord	accord	NOUN	NOUN__Gender=Masc|Number=Sing	ROOT	Accord	NOUN	Confédération
Entre	entre	ADP	ADP___	case	Confédération	PROPN	–
La	le	DET	DET__Definite=Def|Gender=Fem|Number=Sing|PronType=Art	det	Confédération	PROPN	–
Confédération	Confédération	PROPN	PROPN__Gender=Fem|Number=Sing	nmod	Accord	NOUN	entre, la, suisse, Communauté
Suisse	suisse	ADJ	ADJ__Number=Sing	amod	Confédération	PROPN	–
Et	et	CCONJ	CCONJ___	cc	Communauté	NOUN	–
La	le	DET	DET__Definite=Def|Gender=Fem|Number=Sing|PronType=Art	det	Communauté	NOUN	–
Communauté	communauté	NOUN	NOUN__Gender=Fem|Number=Sing	conj	Confédération	PROPN	et, la, économique, européenne, concernant
Économique	économique	ADJ	ADJ__Number=Sing	amod	Communauté	NOUN	–
Européenne	européen	ADJ	ADJ__Gender=Fem|Number=Sing	amod	Communauté	NOUN	–
Concernant	concerner	VERB	VERB__Tense=Pres|VerbForm=Part	acl	Communauté	NOUN	assurance, assurance
l’	l’	DET	DET__Number=Sing|Poss=Yes	det	assurance	NOUN	–
Assurance	assurance	NOUN	NOUN__Gender=Fem|Number=Sing	obj	concernant	VERB	l’, autre
Directe	directe	NUM	NUM__NumType=Card	advmod	autre	ADJ	–
Autre	autre	ADJ	ADJ__Number=Sing	amod	assurance	NOUN	directe
Que	que	SCONJ	SCONJ___	mark	assurance	NOUN	–
l’	l’	DET	DET__Number=Sing|Poss=Yes	det	assurance	NOUN	–
assurance	assurance	NOUN	NOUN__Gender=Fem|Number=Sing	Obj	concernant	VERB	que, l’, vie
Sur	sur	ADP	ADP___	case	vie	NOUN	–
La	le	DET	DET__Definite=Def|Gender=Fem|Number=Sing|PronType=Art	Det	vie	NOUN	–
Vie	vie	NOUN	NOUN__Gender=Fem|Number=Sing	nmod	assurance	NOUN	sur, la

5. Conclusions

This paper has described and discussed the compilation of CHEU-lex, a trilingual parallel and comparable corpus of Swiss, and to a lesser extent EU, legislative texts. The corpus is richly annotated and provides a new resource for legal language and translation research, while also presenting a fine-grained methodology for the annotation of legal texts. Although there are several corpora consisting of EU texts, many are used to perform NLP tasks and do not require specific metadata. CHEU-lex was conceived primarily to study translation and linguistic variation in multilingual institutions. The diachronic perspective is an additional feature to look at legal language drafting and translation across several decades. The legal texts included required ad-hoc solutions and metadata to account for their formal structure in the three languages and their different features. Texts underwent several manual checks to minimise discrepancies during alignment. However, some noise is still present because of few incorrect formatting of tables, rare encoding problems of non-Latin characters due to HTML, and minor mismatches among parallel segments. The syntactic annotation is also less developed than the others and more prone to errors because it was not corrected manually, and the SpaCy library used for this task was not originally trained on legal texts.

Further improvements of the corpus are planned in the future to encompass new agreements between Switzerland and the EU as well as their implementation in the Swiss domestic legislation. In this way, the corpus is intended to track the relations between Switzerland and the EU in terms of legal drafting and translation. The addition of informative text genres (e.g. press releases) is also under consideration to study the popularising of legal content across languages and institutions.

CRediT authorship contribution statement

Annarita Felici: Writing – review & editing, Writing – original draft, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The corpus was compiled with the technical coordination of Prof. Adriano Ferraresi, from the Forlì Campus of the University of Bologna. Special thanks go to Antonio Contarino, Francesco Fernicola, Silvia Mattiuzzi and Silvia Polito for helping in the practical compilation of the corpus. This work was supported by a grant of the Ernest Boninchi Foundation, but the funders had no role in corpus design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.acorp.2025.100151.

References

Baker, M., 1996. Corpus-based translation studies: the challenges that lie ahead. In: Somers, H. (Ed.), Terminology, LSP and translation. Studies in Language Engineering in Honour of Juan C. Sager, pp. 175–186. John Benjamins.
Biber, D., 1993. Representativeness in corpus design. Literary Linguistic Comput. 8 (4), 243–257.
Biber, D., Conrad, S., Reppen, R., 1998. Corpus linguistics: investigating language structure and use. Cambridge University Press.
Biel, Ł., 2014. Lost in the Eurofog: The textual Fit of Translated Law. Peter Lang. https://doi.org/10.3726/978-3-653-03986-3.
Bowker, L., Pearson, J., 2002. Working with specialized language: a practical guide to using corpora. Routledge.
Elminger, D., 2015. *Les Corpus Feuille Fédérale /Bundesblatt / Foglio Fédérale.* V. 1.2. https://archive-ouverte.unige.ch/unige. 80593.
Höfler, S., Piotrowski, M., 2011. Building corpora for the philological study of Swiss legal texts. J. Language Technol. Computat. Linguist. 26 (2), 77–89. https://doi.org/10.21248/jlcl.26.2011.148.
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V., 2014. Sketch Engine. 1, 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Koehn, P., 2005. Europarl: a parallel corpus for statistical machine translation. In: Proceedings of Machine Translation Summit X: Papers, pp. 79–86. https://aclanthology.org/2005.mtsummit-papers.11/.
McEnery, T., Hardie, A., 2012. Corpus linguistics: method, theory and practice. Cambridge University Press.
Mori, L., 2018. Observing eurolects: corpus analysis of linguistic variation in Eu law. In: Studies in Corpus Linguistics, 86. John Benjamins Publishing Company. https://doi.org/10.1075/scl.86.
Scherrer, Y., Nerima, L., Russo, L., Ivanova, M., Wehrli, E., 2014. SwissAdmin: a multilingual tagged parallel corpus of press releases. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 1832–1836. https://aclanthology.org/L14-1602/.
Schmid, H., Laws, F., 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 777–784. https://aclanthology.org/C08-1098/.
Schmid, H., Baroni, M., Zanchetta, E., Stein, A., 2007. The enriched TreeTagger System. Intelligenza Artificiale, Special Issue On NLP Tools for Italian. IV-2, pp. 22–23.
Sinclair, J., 2005. Corpus and text: basic principles. In: Wynne, M. (Ed.), Developing Linguistic Corpora: A Guide to Good Practice. Oxbow Books, pp. 1–16.
Steinberg, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D., 2006. The JRC-Acquis: a multilingual aligned parallel corpus with 20+languages. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, pp. 2141–2147. https://aclanthology.org/L06-1196/.
Vondřička, P., 2014. Aligning parallel texts with InterText. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, pp. 1875–1879. https://aclanthology.org/L14-1258/.
Xiao, R., 2010. How different is translated Chinese from native Chinese? A corpus-based study of translation universals. Int. J. Cor. Ling. 15 (1), 5–35.