Automatically Updated Corpora of EU National Parliaments with Terminology Extraction in Twenty Languages

Marek Blahuš[1], **Ota Mikušek**[1,2]

[1] Lexical Computing, Brno, Czech Republic
[2] Faculty of Informatics, Masaryk University, Brno, Czech Republic
E-mail: firstname.lastname@sketchengine.eu

Abstract

We present a collection of monolingual text corpora derived from the steno protocols of 30 parliamentary chambers across 22 EU member states, covering 20 languages. The corpora are continuously and automatically updated, enabling intralingual and cross-lingual analysis of parliamentary discussions. Each chamber's protocols are regularly downloaded, processed, and transformed into a unified prevertical text format. A terminology extraction grammar is available for each language, allowing the identification of terms specific to each parliament by comparing the parliamentary debates with a general-language reference corpus (or a custom subsection of the debates to the whole body of them). The corpora include timestamps, enabling the observation of trending topics across all European national parliaments within a single platform. Corpus quality depends on the availability and format of the source data, which ranges from simple text files, DOCX, HTML, to XML and JSON (With documented APIs). A monitoring system ensures ongoing compatibility with any format changes. Currently, the corpora consist of over 2.8 billion words and are managed in Sketch Engine.

Keywords: spoken corpora; parliamentary debates; multilingual corpora; terminology extraction; diachronic analysis

1. Introduction

Parliamentary proceedings[1] are text transcriptions of members' speeches and statements and descriptions of the actions undertaken during parliamentary meetings. Their timely availability is essential for achieving greater parliamentary transparency and accountability. The data is usually collected using computer-assisted stenography or speech recognition technology and enhanced with metadata, such as the agenda, the speaker or the votes taken (IPU, 2014).

While there has long been a text corpus consisting of the proceedings of the European Parliament (Koehn, 2005), widely used for analyzing the European Union's political discourse, substantially more effort is required to perform the collection, processing and unified presentation of the debates in the national parliaments of each of the EU member states.

[1] Parliamentary proceedings are also referred to as *official reports*, *parliament minutes*, *steno protocols*, or *shorthand writings*.

Between July 2020 and May 2021, ParlaMint I (Erjavec et al., 2022) was started as the ParlaMint 1.0 (Erjavec et al., 2020) project to fill in this gap, and was later followed by the ParlaMint II project, whose latest version at the time of writing is ParlaMint 5.0 (Erjavec et al., 2025). The ParlaMint II provides over 1 billion words from 29 European countries. Unfortunately, the ParlaMint project has been discontinued, and the ParlaMint corpora only provide data from 2015 to mid-2022.

Another similar project is ParlSpeech (Rauh & Schwalbach, 2020). Compared with ParlaMint, ParlSpeech covers a smaller set of countries (nine, including two non-EU), although its time span is broader (varying between 21 and 32 years). The project ended in 2020 and no new data has been released since then.

In 2023, we published a new tool set for producing parliamentary corpora of EU member states (Mikušek, 2023). This tool set was designed with a focus on continuous and automatic updates and ease of maintenance. We now decided to develop it further and start publishing its outcomes, under the name *ParlaTalk*, in the form of regularly automatically updated corpora accessible to the public online through the Sketch Engine corpus management system (Kilgarriff et al., 2014). At the time of writing, the ParlaTalk corpora consist of over 2.8 billion words and contain steno protocols from 22 EU states in 20 different languages.

In this paper, we describe the improvements and challenges in the curation of the corpora and show how terminology extraction using improved terminology extraction grammars can open new ways of studying the ParlaTalk data.

2. Data Acquisition for ParlaTalk

For each Parliament, whenever possible, we identified two URLs: one for discovering new steno protocols and one for downloading them. When a suitable source of steno protocols could not be found, the respective Parliament was contacted and inquired about data availability (as some data sources are not publicly advertised). We download only from official sources because relying on a third party could introduce inaccuracies, loss of data, or manipulation of the data, as well as add an extra layer of dependency.

We ended up being able to continuously and automatically download, process, and generate corpora from the proceedings of the Parliaments of 22 out of the 27 EU member states. Before being published, each corpus is automatically annotated using Sketch Engine's respective language-specific morphological pipeline to enable better linguistic analysis, including terminology extraction.

A distinctive feature of the ParlaTalk corpora is their focus on continuous and automatic updates. Following the initial build, new documents are periodically queried and added to the corpora to keep them up-to-date. Currently, an updated version of the ParlaTalk corpora gets published every 20 days. Frequent automatic updates limit the possibility of manual data fine-tuning, but are essential for achieving ParlaTalk's mission of delivering recent data in a sustainable way. Still, data only becomes available after it is released by the Parliament and made accessible through its data source. This applies to both new and historical (i.e. retro-digitized) data.

2.1 Data Format

Due to the heterogeneous character of the sources, data is gathered in a large variety of formats, including plain text, XML, HTML, DOCX, XLSX, or JSON. During corpus building, all data is being converted into the unified format of a prevertical file (i.e. a plain text file containing lines of corpus text and structural metadata). An example of a prevertical file can be seen in Figure 1 and illustrates how metadata (in the form of SGML-like markup) is held apart from the actual text (the parliamentary speeches).

```
<doc source_url="https://oireachtas.ie/…/2022-07-14/">
        <speaker name="Deputy Pearse Doherty">
        <p>
        So there will be no savings this year.
        </p>
        </speaker>
        <speaker name="Deputy Michael McGrath">
        <p>
        That is the assessment at this time.
        </p>
        </speaker>
        <speaker name="Deputy Pearse Doherty">
        <p>
        Is it likely that ELS may have to dip into the unallocated
        €2.7 billion at this stage? I ask even though I know that it is early.
        </p>
        </speaker>
</doc>
```

Figure 1: Example of a prevertical file (Parliament of Ireland, shortened, some metadata removed)

2.2 Metadata

In the source data, metadata is encoded in differing ways, sometimes varying even within a single resource due to changes of format in time. We have therefore designed resource-specific heuristics for the extraction of metadata.

There are five common metadata that we strive to acquire across all corpora:

94. *name of the speaker*,
95. *date of the session*,
96. *chamber* (when the parliament is not unicameral),
97. the *URL* from which the steno protocol was downloaded,
98. the *access time* at which the steno protocol was downloaded.

Some additional metadata are specific to particular corpora. For example, the Parliament of Finland provides additional and more fine-grained metadata on *context*, *parliamentary*

group, *speaker's first name*, *speaker's last name*, *speaker's role in the Parliament*, *speech start time*, *speech end time*, *speech type*, and a *direct link to the speech*. We include such extra and fine-grained metadata in our corpora where available, but, for the sake of authenticity, we do not use heuristics to extract additional metadata behind the basic set listed above (e.g., we do not try to break down a non-structured speaker's name into last name, first name and role).

To better adapt to changes in the availability of metadata over time, we introduced automatic detection of structural changes. Newly downloaded data gets scanned for the presence of new elements, using manually curated, source-specific allow lists and deny lists. When an irregularity is discovered, the system issues a warning which can subsequently be addressed by the tool set's maintainer, possibly resulting in more metadata becoming available in the particular corpus in future updates.

Morphological annotation is a specific type of metadata, which we attribute to each word of the acquired texts using Sketch Engine's corresponding morphological-tagging pipeline. Due to language-specific nature of the process, the set of morphological tags differs in each of the twenty supported languages. The minimal set of tags available for each word in every language is the actual *word*, its *part of speech*, and its *lemma*.

2.3 Data Quality

The quality of the data collected is directly dependent on the quality of the source. For example, in the data from the Parliament of Germany, which go as far back as 1949, we discovered 80,803 names of speakers which only appear once in the whole dataset. This is illustrated in Figure 2.

By examining the source data, it becomes evident that this is indeed inaccurate and the cause of the problem is the extremely varying format of the source data, which our heuristics fail to fully comprehend. Even though the data is served through an API, the actual source is apparently OCR'ed scans that have neither been proofread, nor annotated with metadata, forcing us to apply error-prone string-pattern searching to maintain our basic metadata standard.

```
99. (Beifall bei der FDP - Zuruf von der SPD:\nAha! - Klaus Brandner
    [SPD]: Da klatscht von\nder CDA keiner!)
100. (Beifall bei der SPD und dem BÜNDNIS 90/\nDIE GRÜNEN - Friedrich
    Merz [CDU/CSU]:\nDas ist ja Voodoo!)
```

Figure 2: Examples of incorrect speaker's name detection in documents of the German lower chamber (numbers 949 and 967). Due to the stray newline characters (`\n`), speaker gets detected as *Aha! – Klaus Brandner [SPD]* and *DIE GRÜNEN – Friedrich Merz [CDU/CSU]*, respectively. This is because, usually, speaker's name stands at the beginning of a line, followed by a colon; a dash is among the allowed characters in a name, and there is no speaker indicated for *Aha!*.

Such inaccuracies in speaker name identification, however, can often be mitigated by using Sketch Engine's corpus query options. Before studying the speeches by a particular

member of the parliament, it is recommended that the user creates a subcorpus which limits the full corpus only to speeches of that person. This is done by selecting whichever names that person is listed under in the data. Autocomplete and pattern matching come to aid. This solution can sometimes find use even when data quality is good, because no source goes as far as linking two completely different names following a person's name change (e.g., change of surname after marriage).

Poor metadata quality is manifested also in the corpus of the Parliament of Bulgaria. The Parliament provides its proceedings in plain text, using two levels of line breaks: one being the combination of the carriage return and line feed characters, the other an HTML-style line break. Examples of outliers, which were incorrectly classified as speakers, are shown in Figure 3. Speaker detection could be improved by introducing more complex heuristics, which, though, would introduce higher maintenance requirements.

```
101. „ § 89:<br />\r\n1.      3        .<br /> …
102. „ § 7:<br />\r\n )    . 1            . 7       ; …
103. 1.   §1:<br />\n )    . 76      „           …
104.                 ,              ,              : „
                  …
```

Figure 3: Illustration of speaker misclassification in data from the Parliament of Bulgaria. The whole phrase before the first colon was classified as the speaker of the text that follows it, while, in fact, no change of speaker happens in this place and the colon has another use here.

Some countries are known for the high quality of their parliamentary steno protocols, which undergo manual correction before publication. This is the case of the Parliaments of Czechia, Slovakia, and the Netherlands. These Parliaments also apply a transitional period, during which complaints can be raised against already published transcriptions. This means that recent steno protocols may still change for a certain time after they have been made available. For higher stability of ParlaTalk data, we ignore such documents at first and only include them in the corpora once this period for changes has expired.

Another source of good-quality data is the Parliament of Finland. It releases its steno protocols in the form of XLSX sheets,[2] which contain the parliamentary speeches annotated with rich metadata (see Section 2.2), arranged in a tabular structure. A new table is started following each election (i.e. every four years) and the latest table is updated irregularly, every six to twelve months.

To cope with the occasional unreliability of the data sources, as well as with possible technical issues in the unsupervised data acquisition and data processing, we have designed the tool set to be fault-tolerant. Its operations are atomic, and the current status of the processing is being stored in a SQLite database. This ensures that, in case of a system failure, the tool is able to resume its work from the last valid state. To facilitate manual troubleshooting, all tool actions are logged.

[2] Parliament of Finland's data is available at: https://avoindata.eduskunta.fi/#/fi/dataset-search

3. Recent Developments in the ParlaTalk Corpora

Even though the ParlaTalk tool set is in principle autonomous, occasional human intervention is still required. Under normal circumstances, the size of the ParlaTalk corpora grows as new data (recent and sometimes also historical) is being added in the sources, downloaded, processed and included in the continuously updated corpora.

The current state of the data set, as of July 2025, can be seen in Table 1, which lists all the ParlaTalk corpora, along with their size, language, and earliest year covered. There are now 23 corpora for 22 countries (Belgium is represented as a double corpus, see Section 3.2), spanning a total of 20 different languages. It can be observed that, in the passing of time, there has been both an increase and a decrease in the size of the collected data for the individual languages. This is because we try to closely follow any changes implemented in the sources, such as document withdrawals or republishing. In the process, mistakes happen on both our and the remote side, leading to possible cases of duplication which have manifested themselves as a sudden growth in size and were removed over time, resulting in a decrease in the amount of words in the corpora. This can be visible, for example, in the ParlaTalk Ireland corpus.

3.1 Data Downtime

One frequent type of error often encountered during data updates is that the Parliament fails to provide data. In most of the cases, however, this tends to be just a connection timeout or a temporary outage, which can be resolved by trying again at a later time.

Sometimes, though, outages are more serious: From February 12th to March 18th, 2025, the website of the lower chamber of the Parliament of Romania was out of service, possibly due to a hacker attack at the time of a nearing political election. After we confirmed that it was an actual outage and that the problem is not just in the tool or it being blocked access to the servers, we attempted to contact the representatives of the Parliament, but received no response. The situation recovered on its own, more than a month later.

3.2 The Multilingual Parliament of Belgium

A challenging complication, which was not addressed in the original ParlaTalk release, is present in the data supplied by the Parliament of Belgium. In this parliament, both Dutch and French are the languages of communication. In the upper chamber, a translation is always provided in the steno protocol, regardless of the original language. However, in the lower chamber of the Belgian parliament, only the original-language version is recorded. As a result, the corpus is a mixture of sentences in either Dutch or French, which makes it more difficult to use, particularly in relation to terminology extraction, because foreign-language words tend to emerge as false-positives in automatic terminology lists due to their lack of presence in the reference data.

We first considered splitting the corpus into two, each containing only sentences in the same language. This, however, would interrupt the flow of the recorded debates, particularly in cases when one speaker was reacting to another speaker and each used a different language. Another idea was to use two processing pipelines (Dutch and French) to annotate the corpus, switching between them according to the identified language of each sentence. This

Corpus Name	Words in 2023	Words in 2024	Words in 2025	Language	from Year (as of 2025)
Austria	10M	11M	14M	German	2019
Belgium	55M	59M	60M	French/Dutch	2007
Bulgaria	5M	15M	8M	Bulgarian	2022
Czechia	29M	34M	24M	Czech	2010
Denmark	79M	80M	90M	Danish	2007
Estonia	9M	12M	11M	Estonian	2020
Finland	21M	23M	26M	Finnish	2015
France	190M	243M	107M	French	2004
(lower) Germany	125M	131M	286M	German	1949
(lower) Greece	58M	59M	77M	Greek	2015
Hungary	3M	3M	56M	Hungarian	2022
Ireland	41M	121M	45M	English	2022
Italy	16M	23M	106M	Italian	2018
Latvia	-	-	1,001M	Latvian	1937
Netherlands	81M	94M	105M	Dutch	2013
(upper) Poland	20M	20M	20M	Polish	2011
(lower) Portugal	141M	141M	147M	Portuguese	1976
Romania	40M	44M	45M	Romanian	2001
Slovakia	7M	10M	12M	Slovak	2022
(lower) Slovenia	15M	26M	87M	Slovenian	2018
(lower) Spain	67M	69M	443M	Spanish	2019
Sweden	132M	132M	135M	Swedish	1994

Table 1: Current state of the ParlaTalk Corpora. The number of words for each parliament is given in millions (M). Year of the oldest document in the corpus is shown. When only the lower or only the upper chamber is covered, this fact is indicated in the corpus name.

would have provided difficulties in querying the corpus (not all tags would be available for all tokens and some tags would mix values from two different tagging systems) and would not get rid of the duplicate sentences (translations) in the data provided by the the upper chamber.

Eventually, we ended up splitting the original ParlaTalk Belgium corpus into two, one for each Dutch and French, processing each of these two corpora with a single morphological tagger, yet marking texts in the other language with the special tag `[foreign_word]`. This makes it possible to automatically identify and exclude them in corpus searches and terminology extraction by creating an appropriate subcorpus.

4. Terminology Extraction for EU Languages

Finding terms in domain-specific corpora has been a feature of NLP tools for more than a decade (see, e.g., (Aker et al., 2013), (Gojun et al., 2012)). While many such tools have been designed as language-independent, the Sketch Engine corpus management system (Kilgarriff et al., 2014) has pioneered language-aware automatic term extraction by means of language-specific grammars, leading to higher-quality results (Jakubíček et al., 2014).

A terminology extraction grammar is a set of rules which define the lexical structures, typically noun phrases, which should be included in term extraction. An example of such a rule is shown in Figure 4. Obviously, not all sequences of tokens of an applicable lexical structure are terms, but Sketch Engine's existing term extraction algorithm takes care of distinguishing actual terms from mere term candidates.

```
define(`common_noun', `[tag="NC.*"]')
define(`preposition', `[lc="a|al|con|de|del|en|entre|para|por|sin|sobre"]')
define(`adjective', `[tag="A.*" | tag="VMP.*"]')
define(`agree', `$1.gender=$2.gender &
$1.number=$2.number')

*COLLOC "%(1.lemma) %(2.lc) %(3.lc) %(4.lc)"
1:common_noun 2:preposition 3:common_noun 4:adjective &
agree(3, 4)
# example: reducción de ojos rojos
```

Figure 4: Simplified example of a rule from the Spanish term grammar, along with definitions of the used macros. The head noun in position 1 is output as lemma, the noun and adjective in positions 3 and 4 must agree in gender and number. The shown example term means "reduction of red eyes". Reproduced from Blahuš et al. (2023).

Earliest term grammars in Sketch Engine typically had the form of a single part-of-speech-based regular expression (e.g., one or more adjectives followed by a noun was the description of a multi-word term in English). However, it has been shown that grammars, prepared exclusively using the linguistic judgment of a native speaker, are substandard to grammars designed while carefully observing lexical structures common in existing terminological databases.

Therefore, we devised a methodology (Blahuš et al., 2023) for designing terminology extraction grammars through a semi-formal process, mimicking authentic terms found in existing terminology databases, thus coming as close as possible to what is considered a term in the language by actual terminologists, who are the authors of such databases. As a proof of concept, we used this evidence-based approach to develop new-generation terminology extraction grammars for seven languages, all of which happen to be official languages of the European Union.

To enhance the usefulness of the ParlaTalk corpora, we have now used this evidence-based approach to develop terminology extraction grammars for seven more languages, namely Bulgarian, Czech, Greek, Latvian, Lithuanian, Polish and Romanian, thus raising the coverage to 14 out of the 24 official European Union languages. For the remaining ten languages, with the exception of Irish and Maltese (which happen to be absent also from ParlaTalk), older grammars are available in Sketch Engine, which typically match terms of limited structural variety and length, although the principle of terminology extraction remains the same. As a result, terminology extraction is currently possible for 22 out of the 24 European languages, and for all of the 20 languages of the ParlaTalk corpora.

Like in the case of the first seven evidence-based grammars, we relied on IATE (Interactive Terminology for Europe, Zorrilla-Agut & Fontenelle (2019)) as the base for developing the seven new grammars. Because IATE contains terms in all the 24 EU languages, it should be possible to exploit it for developing grammars for all the remaining languages in the future.

In Table 2, we provide an overview of how well the new IATE-based grammars perform compared to the old grammars that are based solely on linguistic judgment. It follows the same format as the analogous table in Blahuš et al. (2023).

Language	**IATE terms**[a]	**Old grammar**		**New grammar**	
Bulgarian	46,416	N/A[b]	N/A	23,585	50.8%
Czech	59,356	29,406	49.5%	42,305	71.3%
Greek	285,880	N/A	N/A	131,700	46.1%
Latvian	51,497	N/A	N/A	27,975	54.3%
Lithuanian	69,181	N/A	N/A	33,643	48.6%
Polish	88,049	13,265	15.1%	45,490	51.7%
Romanian	58,540	N/A	N/A	24,577	42.0%

Table 2: Recall of multi-word terms in IATE by old and newly improved term grammars

[a] Numbers for each language are not necessarily based on the same version of IATE, as latest version was used for each grammar's development.

[b] For Bulgarian, Greek, Latvian, Lithuanian and Romanian, there did not exist any terminology extraction grammar before.

For most of the newly supported languages, the resulting term grammar is the first of its kind. For Czech and Polish, where an old-style grammar exists, we observe a significant

increase in coverage of IATE terms. Compared with the first seven languages discussed in Blahuš et al. (2023), where recall was more than 70% for all except for one, we observe somewhat lower recall in the newly developed grammars. This difference might be attributed to the relatively worse morphological annotation available for the new languages, compared to the original set, which consisted of high-resource languages such as English or Spanish. The only low-resource language in that original set was Estonian, which reached the recall of 66.4%, coming close to what we observe for the newly added languages. It is also possible that the recall could still be increased by investing more time and skill into further development of the new grammars (e.g., not all of them have been developed in cooperation with a native speaker).

5. Application of Terminology Extraction on ParlaTalk Corpora

In this chapter, we will demonstrate how terminology extraction from the ParlaTalk corpora can be used to discover topics discussed in the parliaments at various times, as well as to identify favorite expressions and figures of speech particular to individual politicians. The corpora can also become sources of automatically drafted glossaries which cover both the technical terminology of parliamentary debates, as well as topics typical of debates in a selected time frame. Due to the parallel nature of the ParlaTalk corpora, cross-lingual analyses are also possible to an extent.

Terminology extraction is being done by locating term candidates (i.e. single-words and multi-word expressions, acquired by applying the rules that make up a term grammar) in both a focus corpus and a reference corpus, followed by contrasting their relative frequencies. Words or phrases with a significantly higher relative frequency in the focus corpus than in the reference corpus are most likely to be terms. The lists of single-words and multi-word terms in Sketch Engine are ordered by a score called Simple maths (Kilgarriff, 2009), which includes a variable which allows the user to focus on higher or lower frequency words.

Typically, a large general corpus of the same language is being used as reference corpus (typically having the size of billions of words). This solution, though, has two shortcomings: First, the terms acquired in this way are more typical of the *genre* (parliamentary debates) than of the *topics* which are more likely to change in time and more likely what the user is looking for. Second, such large general corpora tend to be several years old and not necessarily balanced. This may skew term extraction results if words are believed to be much more rare in general language than they actually are, simply because the reference corpus does not yet reflect recent developments which may have made some expressions everyday vocabulary very quickly (such as terms related to the COVID-19 pandemic: four of the default reference corpora for the ParlaTalk languages currently come from 2020 and other four are even older).

5.1 Examples of Intralingual Terminology-Based Analyses

Single-words typical for parliamentary debates, calculated in this way, include annotations of the debates (e.g. “applause”, “interrupting shout”[3]), some universal vocabulary typical of (parliamentary) debates (e.g. “chairman”, “amendment”), as well as proper names (of politicians and parties).

[3] This and all the following examples of terms in this chapter are English translations of the actual terms from the respective national languages, for the sake of understandability.

More appealing results are acquired by contrasting a part of a ParlaTalk corpus (e.g. all the debates from a single year, which, in Sketch Engine, users can define as a subcorpus based on text types, i.e. metadata) against the whole of that same corpus. This filters out all the recurrent terms and reveals those that were typical of the selected timeframe. By asking Sketch Engine to show a concordance of the term in the focus corpus, one can for each such term quickly find out the circumstances which made it a trend at that time.

In ParlaTalk Czechia, terms such as "envelope" and "foreign voter" indicate that "introduction of postal voting" (itself a high-scored term) was the most uncommon topic heavily discussed in that country's parliament in 2024. In Estonia, the government wanting to impose a new tax skyrocketed the terms "sweetened drink" and "beverage tax"; there was hardly a talk about these things in the parliament before 2024. In Italy, the term "differentiated autonomy" rose substantially in use in 2024, although the term was in occasional use in the parliament as early as 2007.

As soon as a term has been identified in this way, its usage trends can be visualized using Sketch Engine's Timeline function invoked from the concordance. This is possible due to the presence of timestamps for all documents in the ParlaTalk corpora and, unlike the Trends function, it works for multi-word terms too. Recognizing these keywords and phrases and their popularity in time can assist linguists in producing glossaries or staying up-to-date with the vocabulary (and agenda) of the national parliaments without having to actually watch or read all the debates.

Intralingual comparisons can be performed also across years within the same country, or across countries if they share a language. By extracting terms from ParlaTalk Austria with ParlaTalk Germany for the same years being used as reference, not only names of local parties and politicians, but also other other outliers can be discovered which have different names in Austria than in Germany (e.g. in Austria the lower house is called "National Council", but in Germany "Federal Diet"; there are also grammatical differences between the countries, such as "Federal Minister *for* vs. *of* the Interior"). Historical insights are possible by contrasting different periods within the same country: In the comparison betwen parliamentary debates of (West) Germany in 1970–1974 and 2020–2024, the zeitgeist of that era is reflected by extracted terms, e.g. the now defunct "DM" (currency) and "(the) Soviet side", or words that have since been replaced by modern alternatives, e.g. "remote reporting bodies" (i.e. telecommunications).

Terminology extraction can also be used to discover single-words and multi-words typical of a particular speaker. By creating a subcorpus of a speaker's contributions to the parliamentary debates and contrasting it with the speeches of all the members of the parliament, we could find out that e.g. former Czech prime minister Andrej Babiš's enjoys addressing the public in his parliamentary speeches (93% of use of "dear fellow citizens" in the corpus is his); that the French Jean-Luc Mélenchon's most distinctive topic in the National Assembly was "consumer debt"; or that George Simion, Romania's parliament member, liked using the phrase "punch in the mouth of the opposition" in 2022–2024 (sometimes seconded by his fellow party members), although it is not his invention as there is documented use of it by other deputies back in 2010–2012.

5.2 Cross-Lingual Terminology-Based Analyses

Thanks to European integration, part of the agenda is nowadays same in the EU member states, which means that even if ParlaTalk corpora are not strictly parallel corpora, they show parallels due to same geopolitical background. Cross-lingual terminology comparison is not directly supported by Sketch Engine (it works only for true parallel corpora), but in time and geopolitics, permitting observations on how same topics trended in time in the individual countries.

As an example, WHO announced "COVID-19" as the name of the new disease on 11 February 2020; the earliest recorded parliamentary use of that word, according to the Publication date metadata, was in the second decade of February in Germany and France; in the third decade of February in Italy, Romania, Austria, Finland and Poland; in the first decade of March in Belgium, Portugal, Latvia, Netherlands, Sweden and Greece; in the second decade of March in Denmark, Slovenia, Spain and Czechia; in the third decade of March in Hungary. It must be noted that this does not necessarily follow only the spread of the disease, but also the speed of acceptance of the new name (earlier uses could be find by looking up names such as "novel coronavirus"), and the exact dates are influenced by the meeting schedules of the individual parliaments. Data for four countries (Bulgaria, Estonia, Ireland and Slovakia) could not be retrieved, as the 2020 steno protocols are currently not available for these countries.

6. Conclusion

The ParlaTalk corpora is a comprehensive, multilingual collection of parliamentary proceedings from 30 chambers of 22 EU member states, representing 20 languages.

The ParlaTalk tool set infrastructure enables continuous and automatic download and processing of parliamentary steno protocols, with minimal human intervention, transforming data sources in different formats, such as plain text, XML, HTML, DOCX, XLSX, or JSON, and with different metadata, into a unified prevertical text format with a minimum shared set of metadata, which is suitable for further linguistic analysis. In the first part of the paper, we have showed that the fact that ParlaTalk's architecture has been designed to be fault-tolerant ensures functionality even in cases of source errors, source format changes, or system failures.

In the second part, we have demonstrated that the ParlaTalk corpora provide an opportunity to observe and study trends in political discourse over time and across countries, and how this can be facilitated by the application of terminology extraction. By using improved language-specific grammars and by applying terminology extraction to carefully selected focus and reference corpora, users can draw various kinds of knowledge on word and phrase usage from the corpora and observe its distribution across time, speakers and countries.

The ParlaTalk corpora, along with the developed terminology extraction grammars, are accessible to the public online through the Sketch Engine corpus management system.

7. Acknowledgements

We would like to thank Mārīte Krupova, Vlasta Ohlídalová and Katarína Petreková for assistance with grammars for particular languages.

8. Future Work

Work to be done in the future includes adding new sources of parliamentary data. Currently, Croatia, Cyprus, Lithuania, Luxembourg and Malta are missing from ParlaTalk. The Parliaments of Malta and Croatia do provide digital steno protocols, but they are currently in a format not suitable for continuous automatic corpus development. In the remote future, the project could possibly be enlarged to cover also non-EU and non-European countries.

There are, and always will be, shortcomings in the quality of the corpus data and metadata, owed to an extent to the heterogeneous and changing nature of the data sources that we rely on. There is space for improvements particularly when the data sources are cooperating, such as the Parliament of Spain, which has recently started publishing its steno protocols also in the XML format, in addition to the HTML used by now.

Last but not least, comparison and possibly some kind of blending with the ParlaMint corpora could be discussed, for the benefit of the user trying to query parliamentary data. If the ParlaMint corpora are processed using the same morphological pipelines, they could be used for terminology extraction using the same terminology grammars which we have described in this paper.

The resulting corpora could be enriched by other methods, such as named-entity recognition, sentiment analysis, or opinion mining.

Software

Aker, A., Paramita, M.L. & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 402–411.

Blahuš, M., Jakubíček, M., Cukr, M., Kovář, V. & Suchomel, V. (2023). Development of Evidence-Based Grammars for Terminology Extraction in OneClick Terms. *Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference*, pp. 650–662.

Erjavec, T., Grigorova, V., Ljubešić, N., Ogrodniczuk, M., Osenova, P., Pančur, A., Rudolf, M. & Simov, K. (2020). Multilingual comparable corpora of parliamentary debates ParlaMint 1.0. URL http://hdl.handle.net/11356/1345. Slovenian language resource repository CLARIN.SI.

Erjavec, T., Kopp, M., Kuzman Pungeršek, T., Ljubešić, N., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M.d.M., Calzada Pérez, M., Cardoso, A., Çöltekin, Ç., Coole, M., Darġis, R., de Libano, R., Depoorter, G., Diwersy, S., Dodé, R., Fernandez, K., Fernández Rei, E., Frontini, F., Garcia, M., García Díaz, N., García Louzao, P., Gavriilidou, M., Gkoumas, D., Grigorov, I., Grigorova, V., Haltrup Hansen, D., Iruskieta, M., Jarlbrink, J., Jelencsik-Mátyus, K., Jongejan, B., Kahusk, N., Kirnbauer, M.,

Kryvenko, A., Ligeti-Nagy, N., Luxardo, G., Magariños, C., Magnusson, M., Marchetti, C., Marx, M., Meden, K., Mendes, A., Mochtak, M., Mölder, M., Montemagni, S., Navarretta, C., Nitoń, B., Norén, F.M., Nwadukwe, A., Ojsteršek, M., Pančur, A., Papavassiliou, V., Pereira, R., Pérez Lago, M., Piperidis, S., Pirker, H., Pisani, M., Pol, H.v.d., Prokopidis, P., Quochi, V., Rayson, P., Regueira, X.L., Rii, A., Rudolf, M., Ruisi, M., Rupnik, P., Schopper, D., Simov, K., Sinikallio, L., Skubic, J., Tungland, L.M., Tuominen, J., van Heusden, R., Varga, Z., Vázquez Abuín, M., Venturi, G., Vidal Miguéns, A., Vider, K., Vivel Couso, A., Vladu, A.I., Wissik, T., Yrjänäinen, V., Zevallos, R. & Fišer, D. (2025). Multilingual comparable corpora of parliamentary debates ParlaMint 5.0. URL http://hdl.handle.net/11356/2004. Slovenian language resource repository CLARIN.SI.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M.C., de Macedo, L.D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M. & Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. URL https://doi.org/10.1007/s10579-021-09574-0.

Gojun, A., Heid, U., Weissbach, B., Loth, C. & Mingers, I. (2012). Adapting and evaluating a generic term extraction tool. In *LREC*. pp. 651–656.

IPU, 2014 (2014). Technological Options for Capturing and Reporting Parliamentary Proceedings. URL https://www.ipu.org/resources/publications/reference/2016-07/technological-options-capturing-and-reporting-parliamentary-proceedings. Document prepared by the United Nations Department of Economic and Social Affairs and the Inter-Parliamentary Union through the Global Centre for ICT in Parliament.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P. & Suchomel, V. (2014). Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 53–56.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, volume 6. University of Liverpool Liverpool, pp. 41–55.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, pp. 7–36.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference*. International Association for Machine Translation, pp. 79–86.

Mikušek, O. (2023). Continuous automatic development of European parliamentary corpora [online]. URL https://is.muni.cz/th/ub78x/. Supervisor: Miloš Jakubíček.

Rauh, C. & Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. URL https://doi.org/10.7910/DVN/L4OAKN.

Zorrilla-Agut, P. & Fontenelle, T. (2019). IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond. *Terminology*, 25(2), pp. 146–174.