Proceedings of ParlaCLARIN III @ LREC2022 pages 25–32
Marseille, 20 June 2022

A French Corpus of Québec's Parliamentary Debates

Pierre André Ménard, Desislava Aleksandrova
Centre de Recherche Informatique de Montréal, Université de Montréal
menardpa@crim.ca, desislava.aleksandrova@umontreal.ca

Abstract

Parliamentary debates offer a window on political stances as well as a repository of linguistic and semantic knowledge. They provide insights and reasons for laws and regulations that impact electors in their everyday life. One such resource is the transcribed debates available online from the *Assemblée Nationale du Québec* (ANQ). This paper describes the effort to convert the online ANQ debates from various HTML formats into a standardized ParlaMint TEI annotated corpus and to enrich it with annotations extracted from related unstructured members and political parties list. The resulting resource includes 88 years of debates over a span of 114 years with more than 33.3 billion words. The addition of linguistic annotations is detailed as well as a quantitative analysis of part-of-speech tags and distribution of utterances across the corpus.

Keywords: French, Québec, ParlaMint, Provincial Parliament

1. Introduction

A critical mass of parliamentary corpora (Erjavec et al., 2021) has been published in recent years in many countries (Andrej Pančur and Erjavec, 2018; Onur Gungor and Çağıl Sönmez, 2018; Steingrímsson et al., 2018; Eide, 2020), helping digital humanities research in fields such as political science (Abercrombie and Batista-Navarro, 2020), sociology (Naderi and Hirst, 2018; Dorte Haltrup Hansen and Offersgaard, 2018), etc. Every one of these corpora also informs linguists and natural language processing experts on multiple phenomenon such as named entities, multi-word expressions, sentiment and emotion expressions, regionalisms, foreign words, to name a few. For some languages, national or provincial parliament corpora are the only large, publicly available textual resource serving as a witness to cultural and linguistic change.

This is the case for the *Assemblée Nationale du Québec* (ANQ), or National Assembly of Quebec in English, the legislative body of the only province with French as the official language in a country with both French and English as official languages. One other province is officially bilingual and the others have English as their official language. The main contribution of this paper is the transformation of ANQ's parliamentary proceedings into a TEI formatted resource which is currently only available for collaborative research, together with the annotations, either extracted from the source data or compiled from other sources.

This article begins with a presentation of the main content source (Section 2) and a detailed account of the acquisition process including processing and XML structure (Section 3). Details on both derived and linguistic annotations (Section 4) are followed by a selection of descriptive statistics (Section 5) to better illustrate the content and potential usage of the corpus.

2. Source Content

The transcriptions of parliamentary debates of the National Assembly of Québec are available as HTML on their website[1]. They are organized in *legislatures*, where a legislature designates the collective mandate of the members of a legislative assembly between two general elections.

Each legislature consists of one or more separate *sessions*, at the discretion of the government. A session designates the period that the Assembly sits, including periods of adjournment. It corresponds to the period of time, within a legislature, that elapses between the convocation of the Assembly and its prorogation or dissolution. The government convenes a new session when it intends to breathe new life into a legislature or to specify the objectives of its mandate. To this day, a legislature has had no more than six sessions and some have lasted a single day.

While the current structure is unicameral, it had a second non-elected high chamber from 1867 to 1963 called the *Conseil Législatif* (Legislative Council). Since then, this second chamber has never been reactivated. The other nine provinces and three territories that make up Canada are also unicameral, while the federal structure is bicameral.

The first provincial legislature of Quebec (not available in the online corpus) was formed in 1867 and had 71 elected members of the National Assembly (MNAs), one for each provincial electoral division. Some electoral divisions may span over a whole administrative region of the province while others, in large cities, are restricted to a single neighbourhood.

The current National Assembly is composed of 125 women and men elected in an electoral division under the first-past-the-post system. The leader of the political party that wins the most seats in the general election normally becomes prime minister. Each elected member, past or present, has an online information

[1] http://www.assnat.qc.ca/en/.

page describing aspects of their political and professional life, as well as their involvement in different organisations.

The earlier debates available online, from 1908 to 1963, were reconstituted from members' notes, assembly summaries, newspaper reports and other sources by a team of historians at the Library of the National Assembly of Quebec (Gallichan, 1988; Gallichan, 2004; Saint-Pierre, 2003). The resulting text is mostly in narrative form with the speaker's name and function in the speech turn followed by their words or actions. Between 1964 and 1989, the transcribed debates were curated in order to remove syntactical and style errors. Word order was modified to better follow syntactic logic. Anglicisms were systematically removed and synonyms were used to avoid repetitions.

Since 1989, the respect of the verbatim of statements requires that only minor spelling or grammatical changes may be made to adapt spoken language to written language (e.g. gender and number agreements). No corrections that modify the style or vocabulary are authorized, even slips of the tongue are transcribed verbatim. As a result, the proceedings of the last three decades contain more examples of spontaneous, unscripted speech. Such speech can be seen in Example 1.

"*Ça, c'est la réalité concrète, M. le Président. Et, quant à la députée, peut-être, quand elle va se relever... Elle a eu le temps de réfléchir. Les discussions qu'elle a eues avec M. Arsenault pour qu'elle continue à voter contre la tenue d'une commission d'enquête, là, à quel endroit... C'était quoi, le deal avec lui, là? C'était quoi, l'entente que vous avez eue pour refuser aux Québécois d'avoir une commission d'enquête...*"

(Free translation)

"*That is the concrete reality, M. President. And, about the member, maybe, when she gets up... She had time to think. The discussions she had with M. Arsenault for her to continue to vote against the holding of a commission of inquiry there, where... What was the deal with him there? What was it, the agreement you had to refuse to Quebecers to have a commission of inquiry...*"

Example 1: Excerpt of an unscripted utterance on February 20th, 2014.

Currently, the transcribers at the ANQ publish a preliminary draft of the proceedings at the end of each day of debates. This temporary version is available in a formatted HTML page. A few days later, a revised and approved version (still in HTML) is made available, along with additional documents such as recordings of the debates (audio and video), order paper and notices (pdf), documents tabled, bills introduced, votes and proceedings detailing each vote (pdf). While the accompanying documents are available in both French and English, the debates in both videos and transcribed versions are only in French. As such, this resource this gives a rare historical view of French spoken in Québec since the start of the previous century, as it can differs from other international or regional variances.

The elements distinguishable on a proceeding's page are as follows (Figure 1):

speech turn: A string of text announcing the author of the utterance that follows. It may or may not be contained in the same HTML tag as that utterance and usually ends with a colon. It identifies the speaker by their name (1), their function (2) or both, or it describes the source of the unidentifiable speech (e.g., voices, one voice, ministers, a speaker, cries, a journalist, etc.) (3, 4).

utterance: A string of text consisting of one or more sentences in one or more paragraphs which follows a speech turn (5).

non-verbal block content: Any non-verbal content formatted in a separate HTML tag (most often, centered and in bold). These elements are written testimonials of either document depositions or announcements of actions and speakers.

non-verbal inline content: Inline strings enclosed in parentheses or <b> tags are occurrences of non-verbal content of one of several types: applause, adjournment, indentation, editor note, reference resolution, written clarification, textual reference.

time: emphasized text in parentheses (6).

pre-formatted content: The contents of HTML tags used consistently across periods <table>, <i>, <acronym>, <a>.

Table 1 provides an overview of the available online content. Some years, especially around the first and second World Wars, are unavailable, which explains the difference between total and spanning period.

Legislature	28
Sessions	78
Total period	88 years
Spanning period	114 years
Debates periods	5,948 days
Members	1,310

Table 1: Overview of available content data from the ANQ source website.

3. Corpus Creation

Converting the online ANQ corpus from HTML to a fully TEI-encoded corpus involved many steps and required a series of design decisions. Since the ANQ corpus is being served online for consultation purposes, there was no available API or any convenient download functionality to rely upon. As such, this section details some particular aspects of the source data and design decisions made during the process.

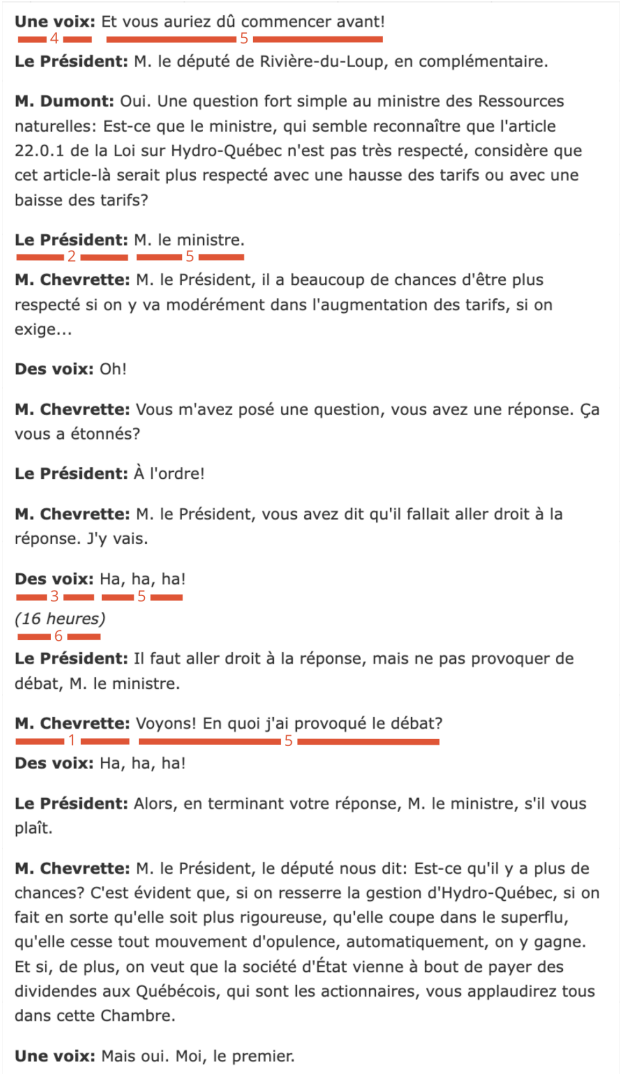


Figure 1: Excerpt of a proceeding's web page from March 12, 1996 with annotated elements.

3.1. Acquisition

Two main sections were targeted on the ANQ website: the proceedings of each sitting and the individual pages of MNAs who took part in them.

The proceedings were crawled semi-automatically, starting from a list of sittings for a given month. These lists are accessed by selecting a session from a legislature (ex: 42th legislature, 1st session) followed by the desired month. The lists were parsed using the Beautiful Soup HTML parser (Richardson, 2007) to obtain the debates' URLs. The links were then followed to retrieve the complete HTML source of the web page. The page was then cached locally for further processing (i.e. text extraction, normalisation, annotations, etc.) and to avoid retrieving it multiple times. Politeness constraints were applied to the crawler to limit the public server load.

Starting from an online list of all MNAs since 1764, the information page of each member was crawled automatically. Current members have HTML-tabbed information page listing their function, biography, indexed interventions in the proceedings, press review, submitted bills, expense reports and contact information. The page's header also details the member's electoral division, the political party they are affiliated with (if not independent) and the role they occupy as a MNA (e.g., opposition's speaker on matters of justice, economy, etc.) The information pages of some members (from earlier legislatures) contain an unstructured summary on a single page which makes it more difficult to parse the information reliably.

To complete the information on MNAs and associate each member with an electoral division, we crawled another section of the ANQ website providing an index of MNAs for each legislature. This was essential to disambiguate the family names used to specify the speaker of an utterance, as they were normalized with fully upper-cased surnames followed by capitalized first name and the electoral division separated with a dash (e.g. SURNAME Forename - Division).

Finally, a list of political parties was parsed semi-automatically from Wikipedia [2] to obtain the official names of the current and historical political organisations with elected members. The acquisitions of members' information pages followed the same overall processing steps (HTML parsing, caching, etc.) as the proceedings.

3.2. Processing

To extract the text from the HTML code and to automatically annotate it, we developed a script using Python, BeautifulSoup, and regular expressions. In the process, we discovered five distinct HTML structures used over various periods of the corpus. The first period spans from 1908 to 1963 and contains the reconstructed proceedings. The second (until 1975) is defined by a lack of formatting around the speech turn segments, making them more difficult to identify and extract. The third period (1975 - 2000) contains some noise as a result of it being automatically digitized (by OCR). The final two periods hold the most recent debates (from October 2000 to the end of 2021) and are relatively clean and well-structured. The text in each of the five periods required a slightly different processing approach.

For each sitting, we parsed its HTML and performed an initial cleanup and correction of erroneous HTML tags (e.g. `<ahref>` instead of `<a href>`). We then attempted to identify and annotate each element of the proceeding (section title, non-verbal element, speech turn, speech, table, etc.) according to its related period. Some elements we could reliably identify based on their HTML encoding alone. Others would have different encoding across periods while remaining visually the same. For example, a section title is always displayed centered and in bold letters, but the 1st period uses a `<b>` tag contained

[2] Wikipedia's list of Quebec's political parties.

in a <p align="center">, while the 5th period employs an <a> tag contained in a <p style="font-weight:bold; text-align: center;">. Finally, to distinguish the speech turn (e.g. Mme Kirkland-Casgrain:) from the actual speech segment in one of the periods, we had to rely on positional clues and punctuation. This is because a single paragraph contained both elements. We made a distinction between indirect (from 1908 to 1960) and direct (from 1963 to present) utterances and speech turn annotations because the indirect ones have the speech turn included in the utterance (Figure 2).

L'honorable M. Gouin (Portneuf) propose, selon l'ordre du jour, que la Chambre se forme en comité général pour prendre en considération un projet de résolution relative au bill 192 concernant le Code municipal de la province de Québec.

Figure 2: Excerpt of a proceeding's web page from March 5, 1915.

Prior to saving the text, the following transformations were applied: corrected common typing errors; standardized spaces, hyphens and line breaks; stripped decorative symbols; reconstructed hyphenated words and concatenated sentences spanning across consecutive paragraphs. The spelling mistakes in the text or in the names of the speakers were not corrected.
In addition to annotating the core elements of a proceeding, we further analysed each speech turn and extracted the available structured data on the speaker. Our custom parser relies on regular expressions and string manipulations to detect and extract, as accurately as possible, the various combinations of surname, forename, division and function contained in a speech turn. Figure 3 illustrates a fraction of the variability of the source, while Listing 1 presents an example of input and output.

La Verge Noire
La Secrétaire adjointe
La présidente Mme Houda-Pepin
Le Président suppléant (M. Cousineau)
L'Orateur suppléant, M. Vautrin
Le Président (M. Ouimet, Marquette)
Son Honneur le lieutenant-gouverneur
M. LE PRESIDENT (M. Gauthier, Berthier)
M. L'Heureux (Gilbert)
M. L'Heureux (président du comité plénier)

forename surname
function division

Figure 3: Examples of speech turns and their components

3.3. Corpus Structure

For the TEI version of the ANQ corpus, the schema of ParlaMint CLARIN was followed and each daily proceeding was encoded in a separate XML file.

```
M. LE PRESIDENT (M. Gauthier, Berthier)
{
    "surname": "Gauthier",
    "forename": "",
    "position": "président",
    "sex": "M",
    "division": "Berthier",
    "type": "person" }
}
```

Listing 1: Speech turn from the corpus followed by the corresponding structured data on the speaker.

The root XML element <teiCorpus> contains the <teiHeader> of the corpus, which in turn contains the metadata for the corpus as a whole. It also includes a list of all identified MNAs with their metadata (forename, surname, division, party, URL) and references to their speeches. The <teiHeader> of the corpus is followed by a series of included <TEI> elements where each of them contains one corpus component (one daily proceeding).

The <teiHeader> of a single TEI file contains document-level metadata: legislative period, session number, date, language, place, as well as a list of identified MNAs whose speeches the file contains, followed by text and tag occurrence statistics.

The body of the document lists in order of appearance all elements of a sitting's proceeding, both verbal and non-verbal. The non-verbal elements were annotated using the tag <note> and the @type attribute was used to categorize them based on their form or function (vote, narrative, summary, comment, time). The speech turn segment, which precedes an utterance and contains the name and/or function of the speaker was annotated as a note of @type "speaker". This choice of annotation allows the inclusion of speech turns (as they appear in the source) but also to signal their non-verbal character. A special case of this rule, where the speech turn is also part of the utterance annotation, is applied to all (reconstructed) proceedings between 1908 and 1963.

Utterances were annotated as <u> and were attributed to speakers with the help of the @who attribute. In a standard TEI file, an utterance contains segments <seg> of text corresponding to paragraphs in the source transcription. In the linguistically annotated TEI file, an utterance contains sentences <s> which contain words <w> and punctuation <pc>. Each word is accompanied by its @lemma and @msd attributes. The components of contracted determiners (e.g. du = de + le) were also included. The @msd attributes details features like the UD part-of-speech tag, gender and number, verb tense and form, pronoun type, depending on the syntactic role of each word.

4. Corpus Annotations

In order to produce a fully annotated TEI corpus, some specialized processing of data sources had to be done to combine extracted and automatically annotated information. This section details these steps, their implications and their limits.

4.1. Speaker Annotation

Linking rich speaker information to each utterance in the ANQ transcriptions required transformation and coupling of multiple sources. The main issue is that MNAs are not referred in a uniform and standardized way in the source content, causing misalignment when combining information. A fuzzy matching algorithm was used when no direct fit was found between the MNAs by division list and the MNAs by legislature and session list. This produced a combined list of each MNA for each session with their associated division.

The association links between MNAs and political parties were then extracted from the historical information pages on MNAs described in Section 3.1. Except for recent members who had their political affiliation described in the header of the page, there are no other standardized expressions of this link. The political party was often mentioned with a contextual template in the form of "...elected member of <party> in <division> in <date>..." with some variations. Some other pages indicated the party as an adjective like "...elected in <year> as a <party>" when the name of the party allowed such adaptation. Then some cases were referred by a short name like the "Bleu" (*blue*) instead of their full name like "Parti Bleu" (*Blue Party*), sometimes even as a single word sentence to denote the affiliation. These texts also contain failed election mentions, which were sometimes written in a similar way as winning elections.

A list of official political party names with corresponding short or adjectival forms was compiled. Every form of this list was used to find exact matches in a description with a known contextual template as shown above. For the description without a match, a fuzzy matching algorithm was applied, falling back to a fuzzy search of standalone party names if no matching context was found. A majority vote was then applied to select the most probable associated party of each MNA, with identified ties to be validated manually. Some of the rare cases of defection where members changed party affiliation after being elected might be missed by this approach.

The combination of these two lists resulted in a speaker reference dataset. It was used jointly with the structured data obtained from parsing the speech turn to identify and attribute utterances to their speakers. When generating the TEI file of a proceeding, a speech turn of multiple people (e.g. M. Galipeault (Bellechasse) et l'honorable M. Gouin (Portneuf)) indicated a summary instead of an utterance. A speech turn of a single person referenced by their function (e.g. L'Orateur, Le Président, etc.) was annotated with an ID combining the name of the function and the numbers of the legislature and session (to allow for further disambiguation in a later version of the corpus). A speech turn of a single person containing their division was unambiguously identified using the speaker reference dataset, since there is a single representative per division per session[3]. A speech turn of a single person containing their name(s) was identified (via the @who attribute) using a fuzzy match against the speaker reference dataset and the list of MNAs for the respective session and legislature. Each utterance tag includes information on the role of the speaker (président, vice-président, lieutenant-gouverneur, greffier, etc.) in the @ana attribute. If no role is indicated in the speech turn segment, the role of "député" is attributed. This current method overgeneralizes and fails to distinguish names of guest speakers from names of MNAs. We consider improving it in future work.

4.2. Linguistic Annotations

In order to produce the linguistically annotated version of the TEI corpus, all the debates were annotated using Trankit 1.0.0 (Nguyen et al., 2021). This tool is a multilingual model based on the Transformer architecture (Vaswani et al., 2017) using the large XLM-roberta model (Conneau et al., 2019) with fine-tuned adapters inserted between the language model's layers, instead of a fully fine-tuned language model. While the model is multilingual and some very rare sentences in the ANQ corpus might be in other languages (like English), only the French annotation pipeline was used.

Each utterance is annotated in a separate CONLL-U formatted file with each single daily debate file generating numerous annotated utterance files. For example, the first session of the 37 legislatures has 200 days of debates and 44,375 utterances. Applying the model on the content of this legislature with a single Titan X 12G GPU took approximately 14 hours, averaging at 52,5 utterances annotated per minute, depending on the length of each utterance. While the process was parallelized, the sum of all annotation times amounted to 28 days for the current version of the corpus. Each parallel process ran a single instance of Trankit over all sittings of a legislature, creating a single CONLL-U file for each day of proceedings.

The model produces universal dependencies (Nivre et al., 2016) part-of-speech labels, morphological features and lemmas used in the annotated TEI. While not included in the current TEI format, it also performed syntactic parsing for future use. Morphosyntactic features (number for nouns and adjectives, person, form and tense for verbs, pronoun gender, etc.) are added in the @msd attribute of the <w> TEI element. The lemma of each word is assigned as the value of the @lemma attribute. Table 2 lists the approximate quantity of each part-of-speech tag in the ANQ corpus,

[3]Barring a few exceptions handled separately.

with nouns (NOUN), determiner (DET) and adposition (ADP) as the top three categories. The "other" (X), particle (PART) and symbol (SYM) categories trail the list with the lowest number of occurrences.

Category	Occurrences (by 1M)
ADJ	1,441
ADP	3,467
ADV	1,987
AUX	1,109
CCONJ	0,677
DET	4,672
INTJ	0,081
NOUN	5,801
NUM	0,400
PART	0,014
PRON	3,253
PROPN	0,705
PUNCT	3,902
SCONJ	0,748
SYM	0,085
VERB	3,381
X	0,025
Total	33,312

Table 2: Part-of-speech categories distribution in the ANQ corpus.

While this tool was trained using mostly international French, like the French version of Wikipedia and other online resources, performance on Québec's French was not an issue as the goal was to give a first overview of the distribution of parts-of-speech in the corpus. A more in-depth inspection of resulting annotations might reveal issues with specific regionalisms or specific idiomatic expressions which sometimes differs from one country to another.

4.3. Non-parliamentary expressions

The ANQ has an official procedure where members can denounce words or expressions which they deem inappropriate in the daily working of the chamber. These expressions can include personal attacks (i.e. idiot, bigot, stupid, liar), unfounded claims or accusations (i.e. racket, collusion, criminal action, stealing surplus, nepotism), associations (i.e. friend of a criminal, friends of the party), unflattering comparisons (i.e. Pontius Pilate, Tartuffe, barking like a wild dog, clown, shylock, eunuch, door mat), among many other types.

The ANQ keeps a record of each time such expressions have been denounced by a member of the national assembly by recording the day, who denounced it, the expression and how many times it occurred.

From an NLP standpoint, this could be used as a seed to perform sentiment or bias analysis. As there are only 393 unique expressions denounced a total of 570 times, this can probably mostly be used as an evaluation set or to bootstrap a few shot algorithms to detect similar expressions like insults, threats, etc.

5. Quantitative Analysis

In order to give an overview of the quantity of data available in the ANQ debates, three statistical representations are shown in Figure 4. The graphics respectively illustrate the distribution of spoken words, utterances and sentence for each year with transcribed debates. It is important to note that the statistics reflected in the graphs only cover direct or reported speech in the transcriptions, thus excluding texts from summaries, topic mentions, title, vote reports, etc. The corpus currently contains a total of 1,27 billion sentences spread across 282,57 millions utterances.

The words and sentences distributions look similar throughout the length of the corpus, while the utterance trend is showing a progressive decrease in number. This seems to indicate that the MNAs would speak for a longer period each time they talk. Other hypothesis can point to more scripted interactions or to less dynamic debates. A more in-depth analysis on the nature and dynamics of verbal interactions would be required to verify these hypothesis.

This new corpus also enables deeper analysis using the metadata compiled on each MNAs by projecting the data on other dimension like sex, region, political party, etc. For instance, the first woman was elected MNA in 1961 among a total of 95 members and was granted the management of a minister the following year.

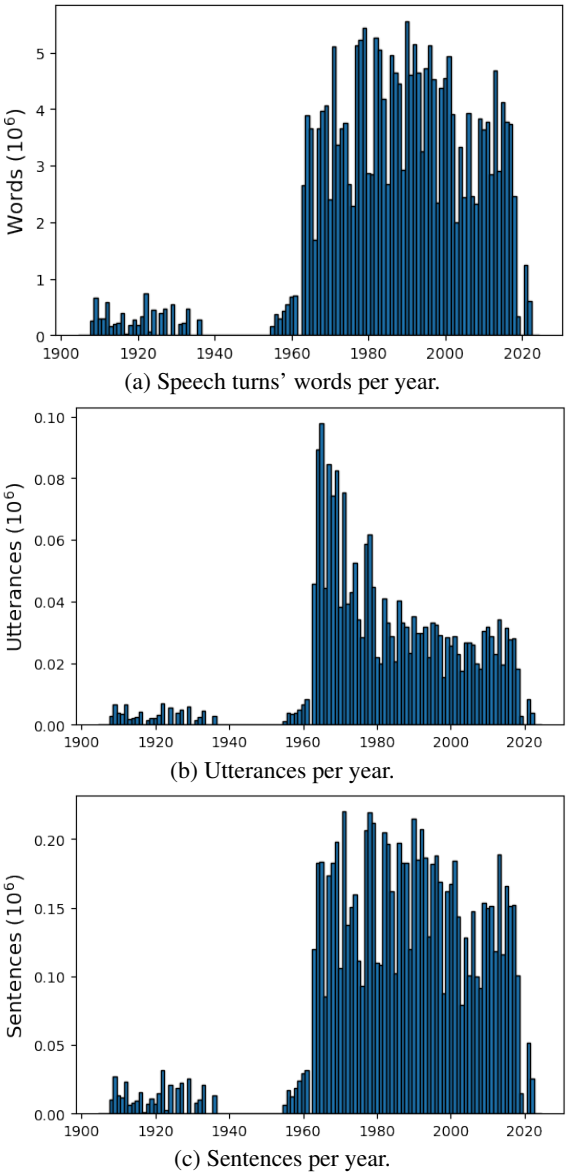
As shown in Figure 5, apart from the missing years between 1960 and 1963, there is a growing proportion of words spoken by female members (in red) compared to males (in blue) since this first occurrence. Clearly, the last two represented years contain much less exchanges as the debates were affected by the global pandemic.

The "*% Elected*" line is projected onto this data to reflected the ratio of male versus female MNAs in proportion to the total number of words spoken by either one for each year. For an easier comparison with the proportion of words spoken by female MNAs, the ratio of elected female MNAs is relative to the top of the bars. As an example, the ratio of elected female MNAs was 44.0% in 2019 (third bar from the right) while the proportion of their spoken words was 37%.

While the representation of women grew steadily from 1% in 1961 to 44% in the last election of 2018, we can see in the same figure that the ratio of transcribed words attributed to female speakers does not follow the same growth rate. This is evidently not a complete analysis by any means, as other variables might influence this ratio like if a MNA is responsible for a minister, if they represent a large city versus a distant region, the attendance rate, etc.

6. Conclusion

This paper presented the creation of a TEI-formatted version of the ANQ online corpus. The process to convert source content into a fully standardized corpus was



(a) Speech turns' words per year.

(b) Utterances per year.

(c) Sentences per year.

Figure 4: Statistical overview of the ANQ corpus.

detailed, as were the data integration tasks needed to enrich it.

Future work would include adding other available but still unprocessed information like topics, named entities, and so on. Other transcriptions will also be added like parliamentary committees' transcriptions which are classified by domain of activity such as education, health, economy, justice, etc. Other levels of annotations could also be added to facilitate the analysis of the discourse's flow such as speakers hesitations, interjections, insults, topics, etc. In addition, complementary annotations could be provided by linking the videos with the transcriptions of the debates and analysing physical communicative phenomenon, like gestures, facial expressions, etc. This would require the integration of tags in Parlamint standard like the <*kinesic*> used in the TEI format of Parla-Clarin, as well as a significant manual or automated annotation effort.

This large dataset of regionalized expressions of a language also enables researchers to train or fine-tune large scale language models to improve natural language processing tools. The annotations and enriched information of the corpus could also help to study the impact of such data on automated tasks like named entities recognition, sentiment analysis, topic modeling, and so on. Improving these tools and resources could support the study of other research hypothesis in academic fields in addition to linguistics and natural language processing. .

7. Acknowledgements

The authors would like to thank the Assemblée Nationale du Québec for their effort, collaboration and access to their data.

8. Bibliographical References

Abercrombie, G. and Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270, jan.

Andrej Pančur, M. S. and Erjavec, T. (2018). SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Dorte Haltrup Hansen, C. N. and Offersgaard, L. (2018). A Pilot Gender Study of the Danish Parliament Corpus. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Eide, S. R. (2020). Anföranden: Annotated and Augmented Parliamentary Debates from Sweden. In *PARLACLARIN*.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L. D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Dargis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Diwersy, S., Luxardo, G., and Rayson, P. (2021). Multilingual

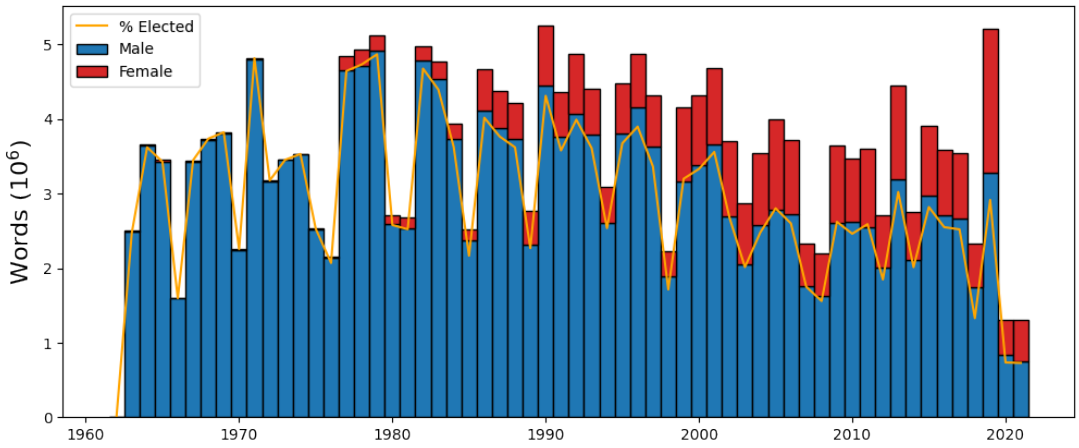


Figure 5: Number of words spoken by sex from 1963 to 2021.

comparable corpora of parliamentary debates ParlaMint 2.1. Slovenian language resource repository CLARIN.SI.

Gallichan, G. (1988). Les débats parlementaires du Québec (1792-1964) et la mémoire des mots. *Papers of The Bibliographical Society of Canada*, 27(1).

Gallichan, G. (2004). Le Parlement « rapaillé »: la méthodologie de la reconstitution des débats. *Les Cahiers des dix*, (58):273–296.

Naderi, N. and Hirst, G. (2018). Automatically labeled data generation for classification of reputation defence strategies. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Nguyen, M. V., Lai, V., Veyseh, A. P. B., and Nguyen, T. H. (2021). Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Onur Gungor, M. T. and Çağıl Sönmez. (2018). A Corpus of Grand National Assembly of Turkish Parliament's Transcripts. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Richardson, L. (2007). Beautiful soup documentation. *April*.

Saint-Pierre, J. (2003). La reconstitution des débats de l'Assemblée législative du Québec, une entreprise gigantesque de rattrapage historique. *Bulletin d'histoire politique*, 11(3):12–22.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Gudnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.