UNIVERSITÉ DE LORRAINE

Jus-TAL

Authors:
Pierre EPRON
Maxime RENARD

Supervisor:
Samuel FEREY
Examiner:
Gael GUIBON

Final report of the Supervised Project for NLP M1

in the

Institut des Sciences du Digital, Management Cognition

April 8, 2024

Contents

List of Figures

List of Tables

Introduction

Constitutional judges play a vital role in modern democracies, particularly in the interpretation and implementation of equality ideals. Since the 1970s, the French Constitutional Council has consistently employed the concept of equality in various forms to evaluate the constitutionality of legislation. Present estimations suggest that approximately half of the Council's judgments incorporate the principle of equality. The Isovote project seeks to investigate the Council's evolving philosophy by examining a previously underutilized source the reports of the Council's debates. Through this research, the objective is to gain insights into how the Council has shaped its philosophy over time.

This tutored project is a component of the Isovote project, aiming to analyze a specific corpus to gain a better understanding of the decision-making process of the Constitutional Council (CC). The objective is to characterize, explain, and comprehend the opinions of each councilor, thereby shedding light on the contributions of the CC to the evolution of the French social state and its assessment of legislative texts pertaining to the fight against inequalities.

Our contributions for this project are:

- Assist the research team in structuring the corpus to ensure its usability for future research projects.
- Analyze and try to detect the nature of direct interactions between advisors, specifically focusing on whether these interactions involve agreement or disagreement.
- Experiment with topic modeling tools on sub-parts of the corpus and specifically parts that mentionned *principe d'égalité*.

We have also submitted an archive file containing all the code and data used for this project.

Chapter 1

Corpus Acquisition

The Constitutional Council (CC) was established by the Constitution of the Fifth Republic, dated October 4, 1958. Its role is to regulate the functioning of public authorities. It is notably responsible for monitoring the conformity of the law with the Constitution. The corpus contains texts which are minutes of CC debates that took place between 1959 and 1995[1]. The particularity of these minutes is that they contain indirect speech about a very specific subject, legal matters, the role of the councillors being to decide whether the matters brought to their attention comply or not with the Constitution. For the purposes of the project, the documents were processed in 4 different ways, summarized in Figure 1.1 and explained below.

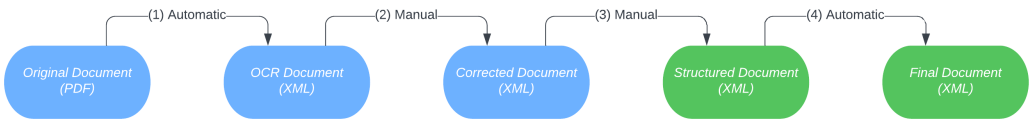

FIGURE 1.1: Diagram of the file transformation process. This project has participated in the stages in green.

1. The original corpus is composed of PDF documents that are processed using Optical Character Recognition (OCR) software that allows them to be transformed into XML files while preserving the content and formatting as best as possible.

2. A sample of the XML files were first corrected by law students during the "Nuit du Droit" of October 4, 2023. The main purpose of this correction was to rectify errors produced during the OCR.

3. A small part of the corrected files was then structured with the help of Professor FEREY. The design of the structure is described below.

4. Structured files are processed by a program that normalizes tags and segments discussions into statements associated with speakers and targets. The design of the program is also described below.

In the following parts, italics are used to refer to the XML identifiers used. These identifiers are in French and will be translated during their first use. For example, the identifier *pv* is used to designate the minutes.

[1] CC minutes

1.1 Schema and Metadata

The various options for the XML schema have been reviewed in the previous bibliography. The Parla-CLARIN (Erjavec and Pančur, 2019) schema was chosen as the ideal solution for the following reasons.

Parla-CLARIN is derived from the Text Encoding Initiative (TEI) schema, which is widely recognized and adopted in the field of digital humanities and linguistic research. TEI provides a comprehensive framework for representing and analyzing textual data, offering robust support for encoding various linguistic features and annotations. It is designed to store a wide variety of linguistic annotations. This flexibility allows for the encoding of different linguistic phenomena, such as named entities, part-of-speech tags, syntactic structures, semantic roles, and discourse markers.

One key advantage of the Parla-CLARIN schema is its ability to directly include speaker metadata within the corpus. When dealing with minutes, it is essential to capture information about the individuals involved in the discussions, such as their names, affiliations, roles, and other relevant attributes. By incorporating speaker metadata into the XML representation, Parla-CLARIN enables efficient retrieval and analysis of specific speakers' contributions, facilitating more nuanced and granular investigations. Parla-CLARIN has been primarily designed to serve scientific goals, making it an ideal choice for transforming minutes into a standardized XML format. As a schema built for research purposes, it provides a solid foundation for scholarly investigations and facilitates collaboration among researchers in the scientific community.

The conception of the metadata structure was supervised by Isabelle Pignonne. We automatized a part of it content:

- List of counselors with their names, birthplaces, roles in the Council, the dates during which they were a member, etc ...
- The name of the proofreader during "Nuit du Droit".
- Identifiers of each decisions and each minutes.

These elements were built following the Parla-CLARIN guidelines[2].

1.2 Structuring of documents

The purpose of this step was primarily to separate the questions, their *rapport* (report), and *discussion* within the *pv* We started by analyzing a sample of *pv*. This sample was made up of all the 11 documents from the year 1982. The choice of the year was made by Pr. FEREY. The structure of the *pv* was found to be dynamic, with variations occurring from one document to another and even within different sections of the same document. This variability presented a significant challenge in automatically identifying and extracting the distinct parts of the *pv*. The absence of a consistent and predefined structure made it difficult to rely solely on pre-defined rules or templates for segmentation. It was therefore agreed that Pr. FEREY would take care of manually structuring the texts. This task simply consists of grouping the paragraphs into tags corresponding to the different parts of a *pv*.

This is a time-consuming task but it can be done in parallel with the proofreading of documents. Indeed, even if the set of documents were corrected during the *Nuit*

[2]Parla-CLARIN guidelines

du Droit event, we found that there were still a good number of errors and that therefore another manual correction step was necessary. The final structure adopted is represented visually in figure 1.2 and in XML in figure 1.3. The decisions made during the realization phase will be explained in the following sections. In the end, we succeeded in finalizing a sample of 32 documents from year 1980 to 1983.

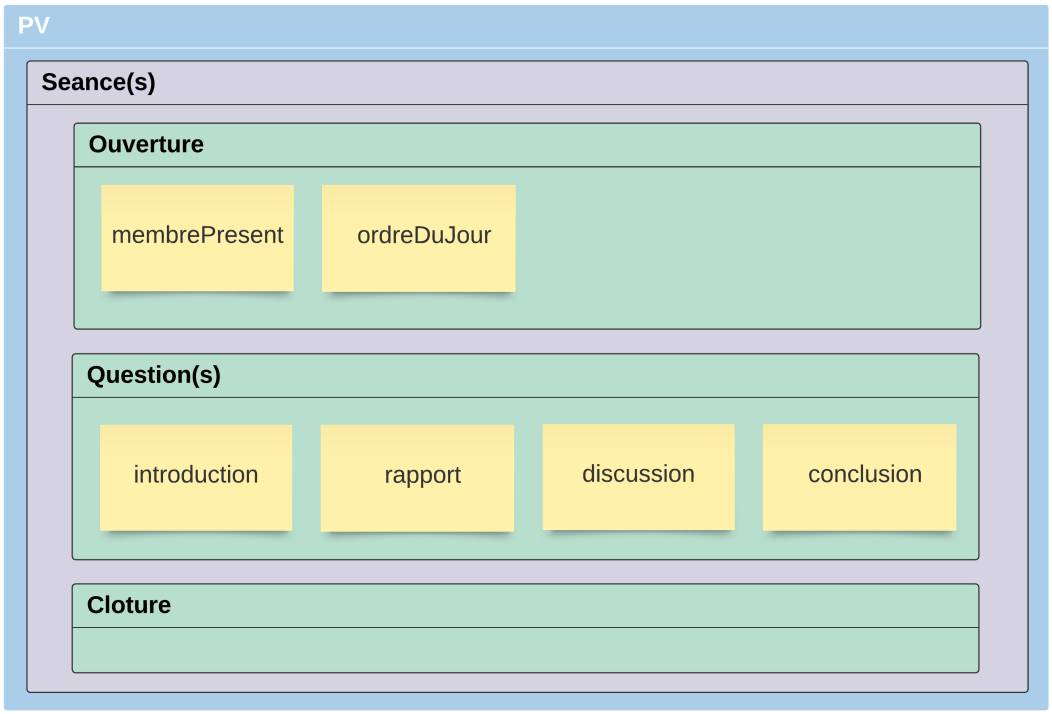

FIGURE 1.2: Visual representation of the structure of a *pv*. Sections ending with (s) can be multiple. All yellow sections can be multiple.

1.2.1 Seance

First, the established structure does not separate the minutes into different *seance* (session). It was assumed that each *pv* consisted of a single session. But very quickly this approach proved to be limited for certain structured documents. For example, the document **PV1982-02-18-23** contains 3 separate *seances*. This happens when all the questions on the agenda of a *seance* are not resolved at the end of this one. In our extended sample (1980 - 1983) out of 32 *pv*, only 4 (12.5%) of them contained more than one *seance*.

A *seance* is defined as a meeting of the members of the CC and identified by the date of the day of this meeting. It is distinguished by an opening (*ouverture*) by the president who generally lists absent members (*membrePresent*) and (re)introduces the agenda (*ordreDuJour*). A *pv* groups together all the consecutive sessions related to the same agenda.

1.2.2 Question

Two distinct types of *questions* can be distinguished: "identified" and "unidentified". Those called "identified" are those that are referenced by the CC. They then gave rise to a decision and there is a full list of these *questions* and their decisions which can be

```
<div1 type="pv" corresp="#pv">
    <div2 type="seance" corresp="#date?(_[a-z])">
        <div3 type="ouverture">
            <div4 type="membrePresent"></div4>
            <div4 type="ordreDuJour"></div4>
        </div3>
        <div3 type="question" corresp="#decision">
            <div4 type="introduction"></div4>
            <div4 type="rapport"></div4>
            <div4 type="discussion"></div4>
            <div4 type="conclusion"></div4>
        </div3>
        <div3 type="cloture"></div3>
    </div2>
</div1>
```

FIGURE 1.3: XML representation of the structure of a *pv*

found on the CC website[3]. Those who are said to be "unidentified" are not referenced by the CC. These are generally referrals deemed inadmissible or *questions* about the internal functioning of the CC. In the final sample, there are only 9 "unidentified" *questions* for a total of 62 (14.5%). There is a good example of a referral deemed inadmissible in the document **PV1982-07-30**.

As explained in 1.2.1, the same *question* can be discussed during several sessions. This rarely happens, only 4 times out of 62 (6.5%), and the most representative example is the one already cited in Section 1.2.1 (**PV1982-02-18-23**); where the same *question* is discussed in three different sessions.

About the content of the *question*, they are divided into two distinct parts. The first part is a report (*rapport*). It is presented by an adviser who lists and categorizes the various referrals to which he begins to respond with one or more arguments. The second part of a *question* consist of a debate between councilors on each referrals (*discussion*. It is punctuated by voting when they does not achieve consensus. In most cases, the *rapport* is presented in its entirety and the discussion follows. However, some councilor prefer to split their *rapport* into parts which are discussed immediately. The question can be therefore structured in a series of *rapport-discussion*. This particular structure remains in the minority in the final sample with only 6 questions out of 67 (9%). The number 67 represents discussion tags and not unique discussions which, as explained above, can be found in different sessions and therefore in different tags.

1.2.3 Content

To represent the content of a structural element (e.g., a report, discussion, etc.), we have defined a set of rules based on the original schema.

All content should be represented using one of the following tags: <u>, <writing>, or <floatingText>.

- The <u> tag should be used to represent discussion elements among councilors.

[3] Les décisions du Conseil Constitutionnel

- The <floatingText> tag should be used to represent external documents added to the transcription.
- The <writing> tag should be used to represent the remaining elements, such as reports and transcript notes.

The text within these tags should be enclosed within <seg> tags. The speaker's identity can be specified using the @who parameter, which can take on several values:

- One or more advisor identifiers should be provided when multiple words are grouped together in the transcription. The identifiers should be separated by a space.
- The shortcut "#president" can be used to refer to the chairman of the board. This allows for time-saving when there is uncertainty about the chairman's identity during the *seance*. The correct identifier will be provided through automation.
- The shortcut "#all" can be used to refer to all the advisors present during the *seance*. No value should be provided when the speaker is not clearly identified or for transcript notes. In such cases, the @who parameter should be omitted.

One of the main difficulties was to clearly define the differences between the types of elements. Even today, the difference between a <u> tag and a <written> tag is not always easy to make and is debatable.

1.2.4 Incident and break

The <incident> tag serves three purposes:

- It represents unusual elements that occur during a *seance*, such as a councilor leaving the room.
- It identifies transcription-related problems that could disrupt the established structure.
- It denotes breaks that occur during sessions.

Elements encompassed by the <incident> tag should be described using a <desc> tag. Regarding the breaks, we initially considered using the <pause> tag. However, upon verification, we found that it cannot contain text or other tags which does not satisfy our requirements. It is important to note that *seance* changes are not considered breaks. A distinguishing factor is that during a break, the *seance* is typically "suspendue" (suspended) while it is "levée" (adjourned) during a *seance* change.

1.2.5 Prospect for improvement

Although the <incident> tag currently serves its purpose, we have identified two issues:

- The inability for the <incident> tag to include a <u> tag or <writing>. For instance, in the case of Vedel leaving the session due to an article in the newspaper Le Monde (PV1982-11-18), being able to identify this part as Vedel's speech could be valuable.

- The need to differentiate between the three use cases.

Votes are currently represented using a tag along with a parameter *@type="vote"*. However, this representation does not allow for the inclusion of the vote results as an XML element. We made attempts to find a solution, but the process is rather complex due to the various ways votes are described. The terminology used for the votes can vary, with some instances employing positive terminology while others use negative terminology. The subject of the vote also influences a lot how the results could be formatted. It has been decided to wait for a larger sample of documents and for the voting process to be further analyzed in order to find a suitable solution. Examples of vote are displayed in 1.4

```
(a) Sont d'avis de déclarer la loi conforme à
la Constitution tous les membres du Conseil à
l'exception de Messieurs GROS, BROUILLET et JOXE.

(b) Ont voté pour le principe des retranchements de mots ou
de membres de phrase, Monsieur le Président, Messieurs
MONNERVILLE, JOXE, GROS, BROUILLET et PERETTI.
Ont voté contre, Messieurs LECOURT, VEDEL et SEGALAT.

(c) La proposition de modification de Monsieur LECOURT est
adoptée par le Conseil.
```

FIGURE 1.4: Example of vote formulation

1.3 Automatic Processing

Despite the challenges posed by the overall structure, an encouraging observation was made regarding the speech interventions of the council members. The majority of these interventions were consistently introduced by the underlined name of the respective council member. Examples of this pattern can be found in Figure 1.5.

```
<p>
    <u>Le Président</u> constate que ...
/p>

<p>
    <u>Monsieur SEGALAT</u>. Monsieur LECOURT, semble-t-il ...
</p>
```

FIGURE 1.5: Examples of speaker introduction in orginal files.

This consistent formatting convention provided a distinctive pattern that could be leveraged for the segmentation process. Taking advantage of this specificity, we designed a program to segment the discussions into utterances using simple regular expressions.

The program analyzed the structured files and applied the designed regular expressions to identify and isolate the speech interventions. These interventions

were then segmented into individual utterances, associating them with the respective council member. This straightforward segmentation technique has significantly reduced the time required to finalize a document, on average by one hour. This achievement represents the most substantial time-saving we achieved through automation.

In addition, we explored the feasibility of employing a Named Entity Recognition model to identify speakers and targets. However, this approach posed several challenges, including the need to differentiate councilors from other individuals. Furthermore, it would have been necessary to establish a method for distinguishing speakers from targets. Ultimately, we would likely have relied on a regular expression (regex) as a potential solution. Consequently, we promptly made the decision to prioritize the simplest approach.

It is important to note that while the segmentation based on underlined names provided a promising starting point, it may not capture all speech interventions. Variations in formatting, missing or inconsistent underlines, or other atypical occurrences might result in incomplete or inaccurate segmentation. Later on the section 2 we provide an estimation of the amount of errors coming from this process.

1.3.1 Targets and Speakers identification

We search for councillors using their family names as they are always referred to this way, except for presidents and rapporteurs, who are frequently addressed using their respective titles.

We generate a list of the speakers that should theoretically be present during the session using the dates during which they sat in the Council. This considerably reduces the list of councillors to search for.

Obtaining speakers for our analysis is a straightforward task, as they are marked with underline tags, allowing us to easily locate them. The only challenge arises when dealing with utterances where the speaker is not explicitly mentioned. However, we address this issue during the document structuring process by manually assigning the appropriate speaker to such utterances.

On the other hand, identifying the targets is more complex. To prevent confusion with councilors' names and unrelated words, we rely on capitalization as a criterion. However, two specific terms, *Président* and *rapporteur* consistently pose challenges. The majority of errors are associated with *Président* since the Council often discusses matters involving other presidents, such as the *président de l'assemblée* or *président du conseil*. Our approach attempts to strike a balance between minimizing these errors and ensuring that we capture mentions of the actual president of the Council. We use the following regular expression to capture *Président*: ([Pp]r[ée]sident)
W(?!d[eu']) which does not catch mentions like *président du sénat* or président de la république.

1.3.2 Other tools

Furthermore, the software carries out various additional tasks, which include:

- Normalizing encoding: The software addresses the presence of uncommon characters that may have been introduced due to configuration issues. For instance, it replaces occurrences of ” with ".
- Hyphen removal: A regular expression is employed to eliminate hyphens from the text.

- Customization of metadata: The software allows for the customization of metadata based on the document's requirements, such as titles and proofreaders.
- Tag standardization: The software ensures consistency in tag usage, substituting <p> with <seg> where appropriate.
- Verification of structural and identifier errors: The software checks for any inconsistencies or errors in the document's structure or identifiers.

1.4 Current state of the corpus

Our work was focused exclusively on the corrected versions of the documents due to the substantial number of errors present in the raw data. This restriction was necessary to ensure the reliability and accuracy of our analysis.

However, we encountered delays in the correction and finalization process, surpassing our initial time estimates. As of now, 39.62% (227 documents) and 5.58% (32 documents) are respectively corrected and finalized out of the total 573 documents comprising the complete corpus (figure 1.6).

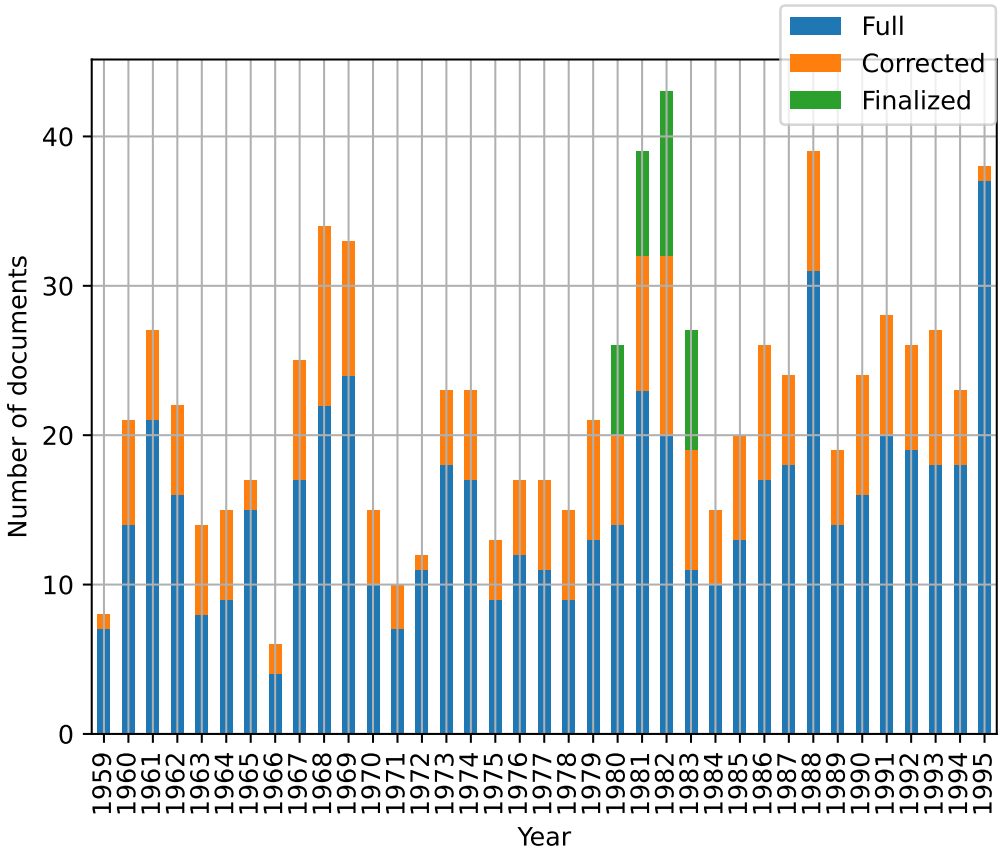

FIGURE 1.6: Number of documents per corpus per year. *Full* designates all the original corpus. *Corrected* designates the part of the corpus proofread during the *Nuit de droit* event. *Finalized* designates the part of the corpus that as been processed by our program.

Chapter 2

Discussion Analysis

The objective defined after the bibliography phase was to annotate the nature and structure of the arguments within a *discussion* sample of the corpus. However, we greatly underestimated the work required for the XML structuring stage described above. We therefore had to find a solution to simplify the task while remaining as close as possible with the original objective. To do this, we took advantage of the fact that we identify the speakers and the possible targets of each utterance.

2.1 Speaker-Target Relations

The main idea is to classify the relationships between an utterance speaker and the different targets in the same utterance. The goal is to be able to quantify the number of times a peer counselor agrees or disagrees. A more formal definition of the task would be: For an utterance U, we have a set of relations $R = (S, T_i)$ where S represents the speaker and $T = \{T_1, T_1, ..., T_n\}$ the set of targets present in the utterance.

This approach obviously has several shortcomings. It only captures relationships when the target is explicitly mentioned in the statement. An utterance in which the speaker simply says that he disagree with an argument without specifying who the argument comes from is not taken into account. To capture this kind of relationship, it would be necessary to consider relationships between the different utterances, which requires a much more complex annotation and has therefore been discarded at this stage of the project.

It is unable to manage complex relationships. The discussions of the CC are very political and the councilor multiply rhetorical figure to express their disagreement while remaining "polite". We therefore frequently have sentences like: "I agree with much of what you say, but I think you are wrong about ...". In this kind of case, the speaker agrees with part of the target's arguments and disagrees with a specific argument. To be able to better classify this type of relationship would require a much more complex approach that identifies the structure of the argument, as originally imagined and finally abandoned.

2.2 Annotation Process

In order to achieve this task, we have annotated a set of utterance extracted from the discussions of the finalized *pv*. This represents a total of 569 utterances but only 243 (43.7%) of them contained at least one relationship. We have annotated the relationships according to 4 labels: *POSITIVE*, *NEGATIVE*, *NEUTRAL*, *CIVILITY*.

POSITIVE and *NEGATIVE* labels correspond respectively to "agree" and "disagree". When we started the annotation, we decided to use very generic terms which did not turn out to be representative anymore.

NEUTRAL and *CIVILITY* labels correspond to relations that are neither agreement nor disagreement. We introduced the *CIVILITY* label because a lot of relationships are about a thanks or something similar. Most of the time the president of the CC thanks the rapporteur for his report. Other advisors do it too. It also happens that a speaker thanks one of the targets for the clarity of his explanation or other. During our first experiments, we quickly noticed that being able to differentiate *CIVILITY* from other *NEUTRAL* could be useful in the analysis of the results. These are often confused with an agreement type relationship.

We have also introduced a *NULL* tag that is used to categorize incorrect relationships that come from potential errors produced when detecting targets.

We carried out the annotation independently from one another using the software INCEpTION[1] and the curation was done with Pr. Ferey. The inter-annotator agreement score between the two annotators is approximately 0.77 (Cohen's K). Most of the differences observed during curation come from cases similar to those described in above. A concrete example : "Sur la demande d'avis, Monsieur VEDEL est en plein accord avec la rédaction proposée mais il tient à indiquer au rapporteur qu'il ne partage pas son opinion quant au fond de la question" which can be translated by "On the request for an opinion, Mr VEDEL is in full agreement with the proposed wording but he wishes to indicate to the rapporteur that he does not share his opinion on the substance of the question". Mr. Vedel initially claims to agree as a matter of politeness, but it is apparent that his true intention is to express his disagreement. It was decided during the curation to consider this kind of example as disagreements. This choice was made with our expected results in mind. Indeed, even if it seems more difficult to design a model capable of identifying this kind of nuance, it is more relevant to consider such a formulation as disagreements when using the results.

The distribution of labels in the curated corpus can be found in the figure 2.1. There is a fairly large majority of *POSITIF* relationships (nearly 43%) which is expected. Since the goal of the discussions is for the councilors to reach an agreement, it seems normal that there are ultimately more agreements than disagreements. There is only 3.8% of the relations that are *NULL*. This is an acceptable percentage for practical use.

Label	Count	Count %
POSITIF	134	42.8
NEGATIF	70	23.4
NEUTRAL	65	20.8
CIVILITY	32	10.2
NULL	12	3.8
Total	313	

TABLE 2.1: Distribution of labels in the annotated sample.

A full example of an annotated utterance can be found in figure 2.1. We kept the XML code for the annotation because some utterances do not directly mention the speaker, as explained in section 1, so it seemed easier to link the tags together. It also facilitates the reverse conversion, i.e. adding the annotate relation to the base XML.

[1] INCEpTION website

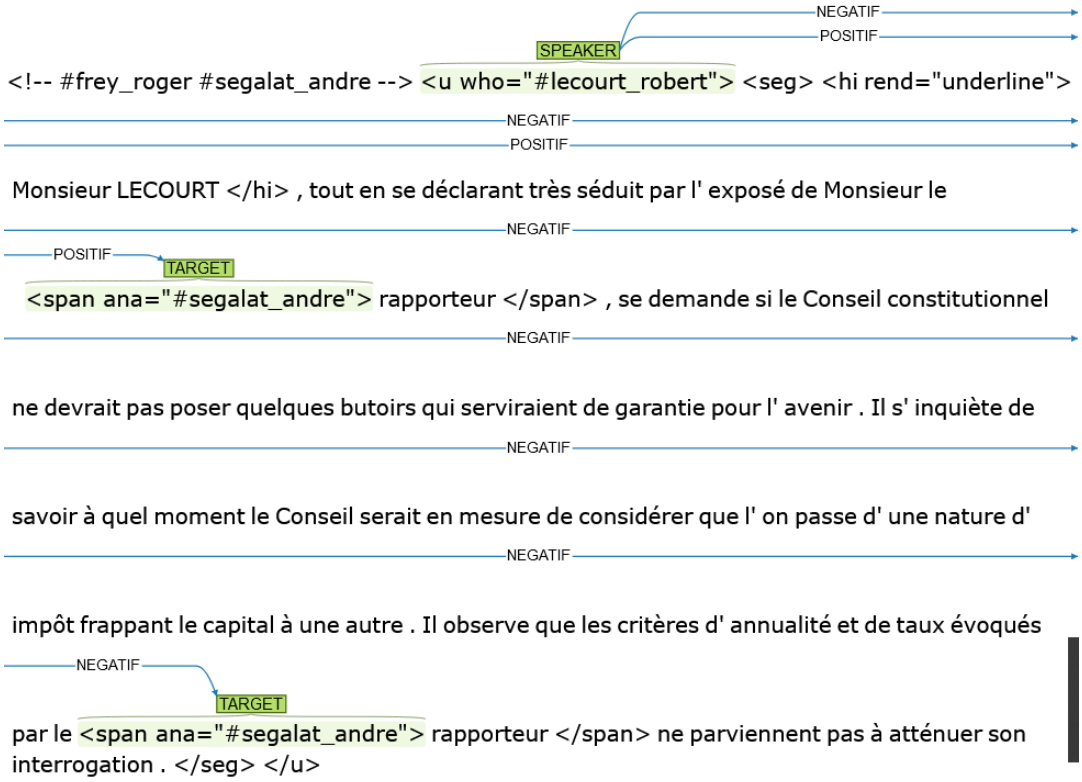

FIGURE 2.1: Screen of an annotated utterance in inception.

2.3 Model

In our specific scenario, we encounter several challenges that require careful consideration. Firstly, we possess a limited set of annotated relations, necessitating a solution that can accommodate this constraint. Additionally, the contexts we encounter are often complex, with the potential to contain multiple relations, each associated with distinct labels. This complexity further emphasizes the need for an adaptable approach that can accurately handle intricate relationships within the text. Furthermore, we operate in a French language context, which imposes restrictions on the choice of available models. Our choice was therefore focused on a NLI type model already trained.

2.4 NLI

NLI (Natural Language Inference) models are designed to tackle the task of determining the logical relationship between a pair of sentences, typically consisting of a premise and a hypothesis. The goal of NLI models is to classify the relationship between the two sentences as entailment, contradiction, or neutrality. These models leverage various machine learning and deep learning techniques to capture the semantic and syntactic information encoded in the text. They often employ neural network architectures such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models to extract relevant features and learn the patterns that govern sentence relationships. NLI models are trained on large-scale annotated datasets, such as Stanford Natural Language Inference (SNLI) (MacCartney and Manning, 2008) and Multi-Genre Natural Language Inference (MNLI)

(Williams, Nangia, and Bowman, 2018), to develop the ability to make accurate inferences across different domains and languages. The advancements in NLI models have significantly contributed to various applications in natural language processing, including question answering, information retrieval, and sentiment analysis and more recently zero-shot classification.

On the Hugging Face hub, we found a model "BaptisteDoyen/camembert-base-xnli"[2] based on camemBERT and trained on the french part of the XNLI corpus.

2.4.1 CamemBERT

CamemBERT (Martin et al., 2020) is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model specifically designed for the French language. It is a state-of-the-art pre-trained language model developed by the Hugging Face team in collaboration with Inria and Facebook AI Research. CamemBERT is trained on a large-scale corpus of French text and provides deep contextualized representations for French words and sentences. It adopts a transformer-based architecture, consisting of multiple layers of self-attention and feed-forward neural networks, enabling it to capture complex linguistic patterns and dependencies. With its extensive pre-training, CamemBERT has demonstrated impressive performance across various downstream natural language processing tasks, including text classification, named entity recognition, and machine translation.

2.4.2 XNLI

The XNLI (Cross-lingual Natural Language Inference) dataset (Conneau et al., 2018) is a benchmark dataset designed to evaluate the performance of natural language understanding models in a cross-lingual setting. It is an extension of the MNLI (Multi-Genre Natural Language Inference) dataset. XNLI consists of sentence pairs in multiple languages, including English, French, Spanish, German, Russian, Chinese, Arabic, and more. The dataset covers a wide range of genres, such as fiction, government, and telephone conversations, to ensure diversity in language usage. Each sentence pair is annotated with one of three labels: entailment, contradiction, or neutral, indicating the logical relationship between the premise and hypothesis.

2.4.3 Selected Model

The "camembert-base-xnli" model has been finetuned on french part of XNLI using the CamemBERT Sequence Classifier architecture of transformers library (Wolf et al., 2020). This consists of adding a head classification to the camemBERT encoder. The ouput of the encoder is pooled and fed to the head which is composed of a dense linear layer of dimension 768 (input and output) and another linear layer used to project the output in 3 dimensions (number of classes). The inputs used are then a composition of the hypothesis H and the premise P sentences following the pattern "<s>P</s><s>H</s>" where <s> and </s> are sentence boundary tokens. To compute the probability of H being an entailment, a contradiction or a neutrality of P, it use a simple softmax function.

[2] camembert-base-xnli

2.5 Experiments

Experiments were conducted using two different models: "camembert-base-xnli," as described earlier, and "xlm-roberta-large-xnli," which is trained on the entire XNLI dataset and is multilingual in nature. However, "xlm-roberta-large-xnli" consistently performed worse than "camembert-base-xnli" across all our experiments. Therefore, we will not present the results of "xlm-roberta-large-xnli" and instead focus on the various modalities we explored with "camembert-base-xnli".

The experiments initially focused on analyzing utterances from the years 1980, 1981, and 1983. The year 1982 was used as a pilot for testing XML structuring, and some adjustments were required to ensure its compliance and usability. While unintended, this turned out to be beneficial as it helped prevent potential biases similar to overfitting, resulting from experimenting with hypothesis and methods. The evaluation presented below encompasses all the years, including 1982.

2.5.1 Utterance segmentation

Using the entirety of an utterance as a premise proved to be impossible due to certain cases where the same target can have two different types of relationships. The model would predict the same results where we expect different ones. To address this issue, we tested two approaches. The first approach involves retaining only the sentence containing the target as the premise. This approach allows us to avoid truncating the premise to fit the model's input dimension (512 tokens) and ensures consistency across all cases. The second approach involves including the entirety of the subsequent sentences that do not contain any other targets as part of the premise. This method provides additional contextual information. Sometimes, the base sentence alone lacks the necessary information for the model to make accurate predictions. Figure 2.2 illustrates examples of both segmentation techniques. In the sample, the only instances of same target with different relation types after segmentation are *CIVILITY* and *NEUTRAL*. Fortunately, this can be resolved by merging them into the *CIVILITY* category without affecting the final results. However, this solution is not perfect, as more challenging cases could arise, such as POSITIVE and NEGATIVE relations in the same sentence for the same target.

2.5.2 Hypothesis test and classification

In this experiment, we explored two distinct approaches for utilizing an NLI model. The first approach involved testing a single hypothesis, such as "X agrees with Y," and directly utilizing the model's results. In this case, the model's output was interpreted as *POSITIF* for entailment, *NEGATIVE* for contradiction, and *NEUTRAL* for neutrality. The second option was to employ the model as a zero-shot classifier. This approach entailed testing an hypothesis for each class and utilizing the corresponding entailment logits as the class logit. For instance, "X agrees with Y" represented the *POSITIF* class, "X does not agree with Y" represented the *NEGATIVE* class, "X is neutral with Y" represented the *NEUTRAL* class, and even "X thanks Y" could represented the *CIVILITY* class.

2.5.3 Hypothesis

The primary challenge of the experiment lies in identifying the appropriate hypothesis. We have conducted many trials and it is difficult to report them all. Considering

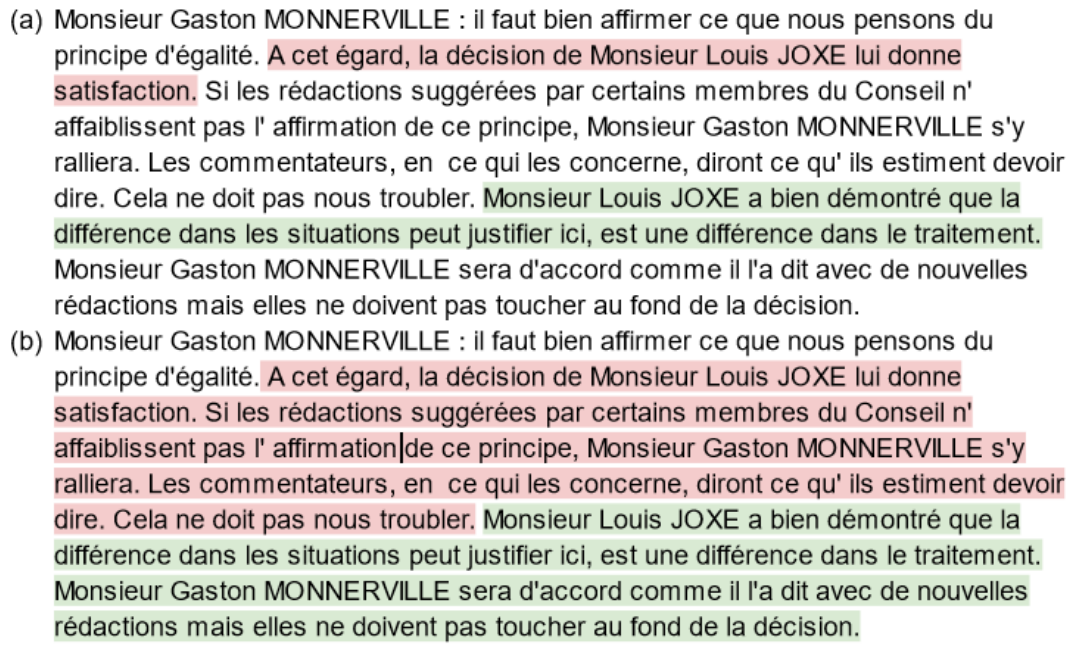

FIGURE 2.2: The two approaches used for segmentation. (a) We keep only the sentence containing the target. (b) We add the following sentences which have no targets.

the *POSITIF* hypothesis "X agrees with Y", we have drawn two conclusions based on all the trials. The phrase "semble d'accord" (seems to agree) has proven to be the most effective wording. Alternative formulations such as "est d'accord" (agrees), "soutient" (supports), or even simply "positive" have shown to be less effective across different template sentences.

Using the speaker's name instead of "X" has demonstrated lower efficiency compared to using "Le locuteur" (The speaker) alone. This observation may seem obvious, as some premises do not mention the speaker. However, it is essential to note that this holds true for premises that do mention the speaker.

2.5.4 Few shot learning

Few-shot classification is a technique utilized to enhance model performance when provided with only a limited number of examples. Generally, this approach proves beneficial for improving NLI (Natural Language Inference) models. In our study, we attempted to employ this technique by utilizing a varying number of examples per class, ranging from 1 to 10. Regrettably, the results were disappointing, as they led to an average decrease in model performance by 0.15. This outcome can likely be attributed to the intricate nature of the premises and their inherent differences.

When examples are provided to the model, it tends to adapt specifically to the provided instances. However, if the examples are not sufficiently representative or if the labels to be predicted exhibit significant variations, the application of few-shot classification can have a negative impact. This observation suggests that our current annotation approach may not be viable in the long term. The defined categories do not adequately capture the content they are intended to represent, leading to suboptimal performance.

2.6 Results

In the evaluation presented in the figure 2.2, we grouped the labels NEUTRAL, CIVILITY and NULL as NEUTRAL so that it best represent the application objective of the project.

We report results from two main types of modalities: **HYP_*** and **CLS_***. It represents respectively the two approaches hypothesis test and classification test described in section 2.5.2.

Generally, HYP outperforms CLS in terms of performance. This observation could be attributed to the inadequacy of hypotheses in accurately representing the expected response. Unlike CLS, which deals with four hypotheses, HYP operates with a single hypothesis, resulting in a relatively lesser impact. This divergence in hypothesis handling may account for the superior performance of HYP over CLS. However, it is important to note that this explanation is merely a personal interpretation and lacks substantial empirical evidence to support its validity.

Configuration	**Weighted AVG**			**POSITIF**			**NEUTRAL**			**NEGATIF**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
HYP_est	0.68	0.67	0.67	0.72	0.74	0.73	0.68	0.62	0.65	0.58	0.63	0.60
CLS_est	0.65	0.64	0.64	0.67	0.72	0.69	0.55	0.64	0.59	0.68	0.55	0.61
HYP_semble	0.72	**0.72**	**0.71**	**0.74**	0.81	**0.77**	0.75	0.68	**0.71**	0.62	0.60	0.61
CLS_semble	0.70	0.67	0.67	0.69	0.80	0.74	0.55	0.69	0.61	0.80	0.51	**0.63**
HYP_chunk	**0.73**	**0.72**	**0.71**	0.71	**0.83**	**0.77**	**0.82**	0.61	0.70	0.61	**0.66**	**0.63**
CLS_chunk	0.70	0.65	0.65	0.70	0.74	0.72	0.46	**0.71**	0.56	**0.84**	0.50	0.62

TABLE 2.2: Compared results of NLI model for different modalities. Precision (P), Recall (R), F-score (F1)

2.6.1 Semble

The modalities ***_est** and ***_semble** are here to demonstrate that hypothesis using "semble" perform better than others. We use ***_est** as comparison because it also has correct performance. The results indicate that all the metrics of ***_semble** show slight improvement compared to ***_est**, except for the recall of *NEGATIF*. The introduction of uncertainty in the hypothesis enhances the model's overall performance, except in accurately identifying all instances of disagreement. Nevertheless, it is important to note that the observed differences in the metrics and the number of relationships are not statistically significant enough to draw definitive conclusions.

Hypothesis for the two modalities are:

- ***_est**
  - POSITIF: "Le locuteur est d'accord avec Y"
  - NEGATIF: "Le locuteur est en désaccord avec Y"
  - NEUTRAL: "Le locuteur est neutre avec Y.
  - CIVILITY: "Le locuteur remercie Y".
- ***_semble**
  - POSITIF: "Le locuteur semble d'accord avec Y"
  - NEGATIF: "Le locuteur semble en désaccord avec Y"
  - NEUTRAL: "Le locuteur semble neutre avec Y.

– CIVILITY: "Le locuteur semble remercier Y".

The *POSITIF* hypothesis is employed when conducting a hypothesis test. Initially, we also explored *NEGATIF* hypothesis, but we consistently obtained inferior results. We hypothesize that the overrepresentation of *POSITIF* relationships contributes to this outcome. The model could have a general tendency of classifying sentences as "entailment." The ***_semble** hypothesis are used for the following modalities.

2.6.2 Chunk

The *_chunk modality utilizes the second chunking method described in section 2.5.1, while *_semble and *_est employ the first method. Surprisingly, despite *_chunk appearing to outperform *_semble, the differences observed are marginal, which contradicts our initial expectations. We had anticipated a more pronounced disparity between the two methods, as it was evident during annotation that using a single sentence posed challenges in various cases. There could be several reasons for this divergence. Firstly, the dataset may lack instances where incorporating context would make a noticeable impact on the results. Secondly, extended contextual information alone may not be sufficient for accurate classification in certain cases, as the model might excessively focus on nearby context. Lastly, the extension of context could introduce excessive noise, impeding the model's ability to focus on crucial elements. Consequently, both methods exhibit flaws and are not considered viable solutions for this particular task. A more precise segmentation of the text based on argument structure would be preferable.

2.6.3 HYP_chunk

The results for the best configuration (HYP_chunk) indicate that there are some notable patterns in the confusion matrix.

Regarding errors related to the NEUTRAL label, it is observed that 15 out of 26 instances (57.7%) labeled as *POSITIF* are actually *CIVILITY*. In contrast, no instances of *CIVILITY* are labeled as *NEGATIF*. This finding suggests that the model faces difficulties in distinguishing between *CIVILITY* and *POSITIF*, especially when the speaker compliments the target. Examples of these errors are reported in figure 2.4

Additionally, out of the 12 instances labeled as NULL, 8 (67%) are mislabeled, with 1 being labeled as *POSITIF* and 7 as *NEGATIF*. This mislabeling is a side effect of problems in target detection.

Examining the errors in the *NEGATIF* category, it is evident that they are predominantly tagged as *POSITIF*. Upon analyzing the context of these errors, it becomes apparent that they often correspond to situations described in section 2.2. These cases involve speakers who, as a form of politeness, initially express agreement before explaining their disagreement. Examples of these errors are reported in figure 2.5

2.7 Application

To conclude this second chapter, it is worth discussing the practical application of this model. The primary objective is to enhance the analysis of dialogues by domain experts. For instance, in Table 2.3, an intriguing interaction between Mr. Vedel and Mr. Segalat is observed. Notably, they exhibit frequent interactions throughout the

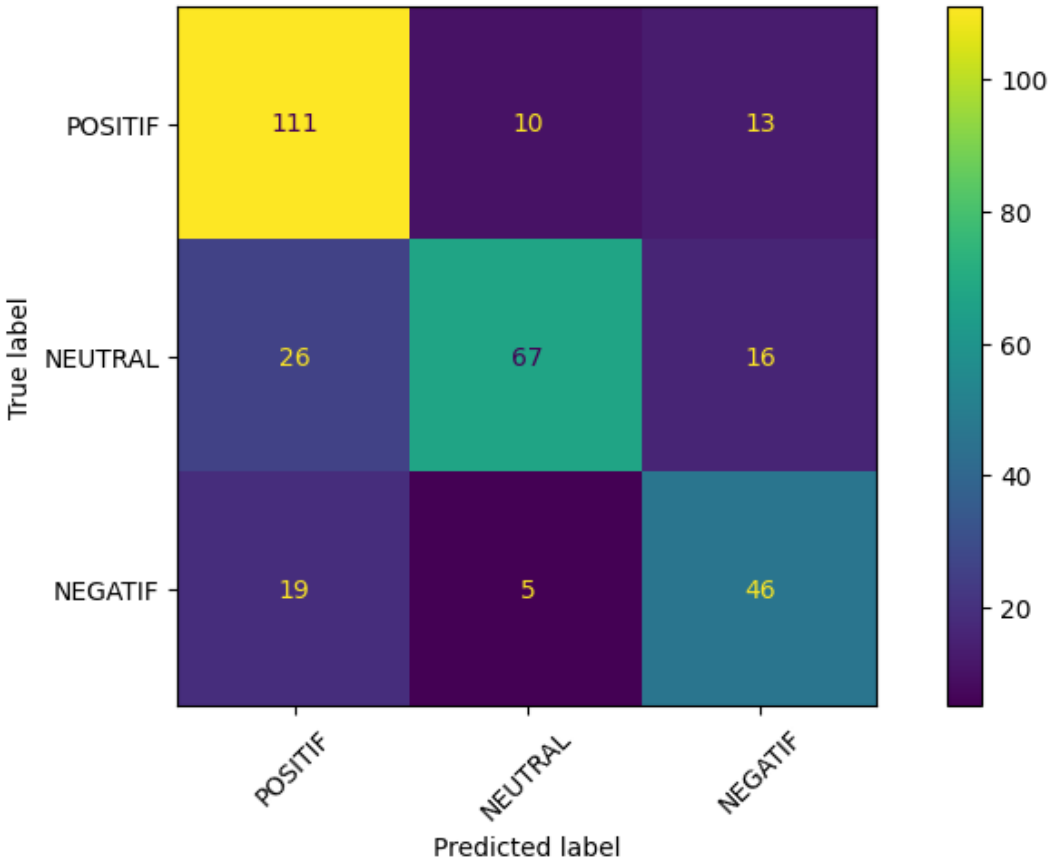

FIGURE 2.3: Confusion matrix for HYP_chunk results.

dialogue. Furthermore, when Mr. Vedel engages in arguments with Mr. Segalat, it is evident that his level of agreement is slightly below 56%. Conversely, Mr. Segalat tends to agree with Mr. Vedel almost 93% of the time. This observation might lead one to speculate that Mr. Vedel exerts some influence over Mr. Segalat. However, it is crucial to acknowledge and address any biases that our process may introduce. Consequently, it is advisable to validate this hypothesis by directly examining the text. Nonetheless, such findings offer valuable leads for further exploration to researchers in the field. A visual representation of the results as a graph is available in Appendix A.

We conducted an evaluation of our automated relationship classification method to assess its viability in an application context. In Table 2.3, the "Prediction" column indicates the agreement ratio between the speaker and the target, as predicted by the **HYP_chunk** configuration. The "delta" column quantifies the absolute difference between the predicted ratio and the gold ratio. In the previous example, Mr. Vedel as speaker and Mr. Segalat as target, a delta of approximately 36 points was observed. On average, across the entire set of relationships, the delta was found to be 25.7 points.

To explore potential correlations between delta and support, we performed a correlation test. This test aimed to determine if the gold support had any significant impact on the delta value. The resulting correlation coefficients were 0.038, 0.156, and 0.395 for the Pearson, Kendall, and Spearman coefficients, respectively. These results suggest that the support does not have a significant influence on the difference between the agreement ratios.

(a) Le Président remercie Monsieur SEGALAT qui a fourni à la fois un travail de fourmi et un travail de géant et qui a exposé avec beaucoup de clarté une question délicate.

(b) Monsieur le Président remercie Monsieur le Rapporteur pour son exposé particulièrement vivant et déclare la discussion générale ouverte.

(c) Monsieur le Président remercie Monsieur LECOURT pour son remarquable rapport.

FIGURE 2.4: Example of *CIVILITY* tagged as *POSITIF*

(a) En ce qui concerne l'article 3, bien qu'ayant suivi avec attention Monsieur le rapporteur, il ne peut concevoir que l'on méprise tout le droit de propriété en imposant des personnes sur des biens qui ne leur appartiennent pas.

(b) Monsieur le President ne méconnaît pas la finesse de l'observation psychologique de Monsieur BROUILLET mais il se déclare également frappé par l'imprécision de la loi en matière de cumul.

(c) Monsieur LECOURT est d'accord sur ce dernier point, mais il n'est pas convaincu par la dialectique habile de Monsieur SEGALAT selon lequel, devant le juge, la règle est identique à l'exception, le juge ayant pour mission de toujours vérifier la condition de réciprocité.

FIGURE 2.5: Example of *NEGATIF* tagged as *POSITIF*

Based on the findings presented in this section, it is evident that employing our solution in practical cases is associated with a considerable margin of error. This outcome is to be expected considering the raw results obtained and discussed in 2.6. However, addressing this issue requires a more sophisticated approach to document analysis. Our current analysis methodology proves to be overly simplistic when faced with the intricate argumentative structures and figurative language employed in CC debates. Therefore, it is imperative to enhance our analytical approach to better capture the complexity inherent in these discussions.

Speaker	**Target**	**Gold**		**Prediction**		**Delta**
		Ag. ratio	***Support***	***Ag. ratio***	***Support***	
segalat	vedel	92.86	14	81.25	16.0	11.61
jozeau-marigne	vedel	80.00	5	100.00	3.0	20.00
vedel	lecourt	71.43	7	66.67	6.0	4.76
segalat	lecourt	71.43	7	80.00	5.0	8.57
gros	lecourt	66.67	9	62.50	8.0	4.17
lecourt	vedel	60.00	5	100.00	4.0	40.00
gros	vedel	57.14	7	80.00	5.0	22.86
vedel	segalat	56.25	16	92.31	13.0	36.06
vedel	gros	37.50	8	100.00	3.0	62.50
lecourt	segalat	33.33	9	100.00	5.0	66.67
gros	segalat	22.22	9	62.50	8.0	40.28

TABLE 2.3: Percentage of agreement (Ag. ratio) between speaker and target. Gold column refer to annotated data. Prediction column refer to predicted data with HYP_chunk. Delta is the absolute difference between Gold and Predicted agreements. The table is filtered to display only peers with a gold support >4.

Chapter 3

Topic Modeling

We attempted several approaches to topic modeling, first, using general unsupervised approaches, with BertTopic, then focusing on *principe d'égalité*. The analysis that follow are conducted on the corrected corpus unless otherwise specified.

3.1 Unsupervized approaches

We attempted to use BertTopic[1] (Grootendorst, 2022) to generate clusters of topics or keywords from the documents. BERTopic is an approach that utilizes three main steps to generate topic representations from a given set of documents:

1. Embedding Generation:

   In the first step, each document in the corpus is transformed into its corresponding embedding representation using a pre-trained language model. Specifically, BERTopic employs a language model that has been pre-trained on a large amount of text data, such as BERT (Bidirectional Encoder Representations from Transformers). This embedding process converts the text into numerical vectors that capture the semantic meaning of the documents.

2. Dimensionality Reduction:

   After obtaining the embeddings for each document, BERTopic aims to optimize the subsequent clustering process by reducing the dimensionality of the embeddings. High-dimensional embeddings can be computationally expensive and may introduce noise to the clustering. Therefore, a dimensionality reduction technique, such as Principal Component Analysis (PCA) or t-SNE (t-Distributed Stochastic Neighbor Embedding), is applied to transform the embeddings into a lower-dimensional space while preserving the essential structure and patterns within the data.

3. Clustering and Topic Extraction:

   Once the dimensionality of the embeddings has been reduced, the next step is to cluster the documents based on their similarity. Clustering groups similar documents together, creating clusters that represent different topics present in the corpus. BERTopic employs a clustering algorithm, such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), to perform this task effectively.

Finally, from the clusters of documents, BERTopic extracts topic representations using a custom class-based variation of TF-IDF (Term Frequency-Inverse Document

[1] BERTopic

Frequency). TF-IDF is a common technique used to determine the importance of terms in a document collection. In BERTopic, this class-based variation of TF-IDF calculates the relevance and significance of terms within each topic cluster, allowing for the identification and extraction of representative keywords and phrases that best describe each topic.

We applied BERTopic on multiple levels of the documents:

- by documents (using the corrected corpus)
- by questions (using the finalized documents)
- by reports and discussions (using the finalized documents)
- by utterances (using the finalized documents)
- by p/seg (using the finalized documents)

And also tried using multiples parameters:

- with/without stopwords
- using different representation models
- using different embedding models

Unfortunately, we found that we either generates too few clusters/too many (up to 300) to be useful and that often, the content of the clusters don't make sense: they would be filled with determiners, with *monsieur* or words like *rapport* or *Constitution* that are common to all clusters. We observed a similar phenomenon for keywords: we would generate generic keywords that are common all elements.

As seen above, the unsupervized methods fail because of multiple factors:

1. we have too little data (in the case of the finalized corpus), which means we cannot focus on one part of the documents, like questions, reports or discussions.
2. the remaining data is not organized, which means we have to use the entirety of the document, which contains a lot of noise.

Therefore, we decided to focus on the Principle of Equality, which is one of the subjects of the ANR Isovote this supervized project is a part of.

3.2 Principle of Equality

The principle of equality, *principe d'égalité* or *principe de l'égalité* in french, within the context of constitutional law and democratic societies, refers to the fundamental concept that all individuals are entitled to fair and equal treatment under the law. It encompasses the idea that no person or group should be unjustly discriminated against or favored based on arbitrary characteristics such as race, gender, religion, socioeconomic status, or any other protected attributes. The principle of equality aims to ensure that every individual has equal opportunities, rights, and protections within society, regardless of their background or personal circumstances.

To analyse the use of *principe d'égalité* we chose to use the entirety of the pre-structured document as, because they were more of them, they contained significantly more mentions of *principe d'égalité*: 411 for 64 mentions in the finalized corpus. Therefore, the analysis that follow are conducted on the utterances that contain at least one occurrence of *principe d'égalité*, in the corrected corpus.

3.2.1 Distribution of principe d'égalité in the corpus

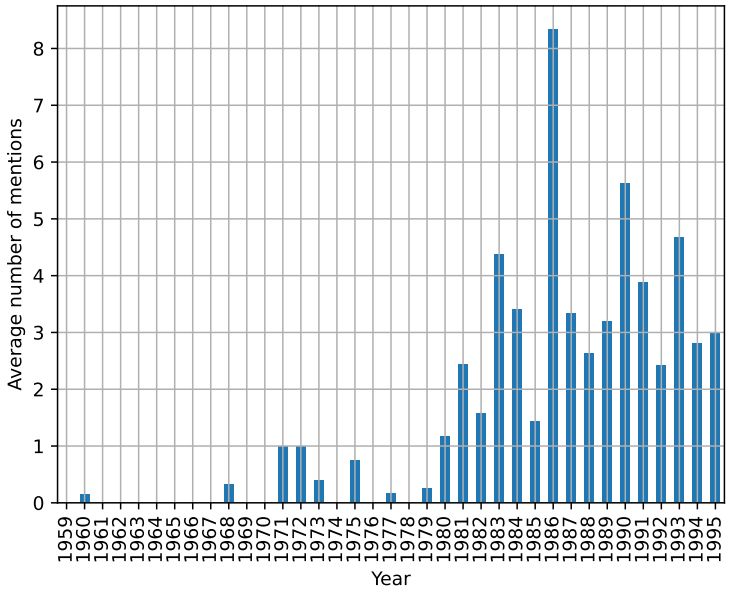

FIGURE 3.1: Distribution of the average number of mentions of *principe d'égalité* per documents per year

As can be seen in Figure 3.1, we observe an overall augmentation of the mention of *principe d'égalité* per document throughout the years, starting from 1981 and with a peak in 1986.

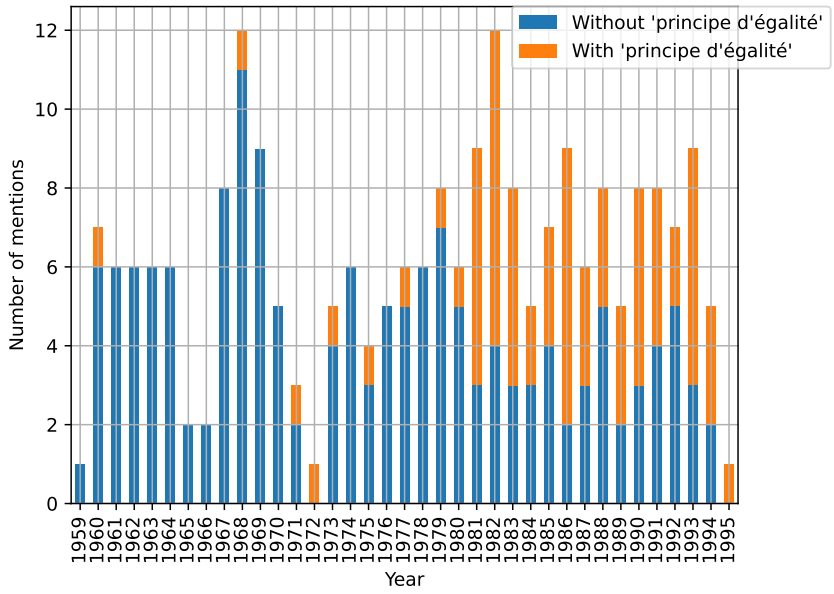

FIGURE 3.2: Distribution of the number of documents mentioning *principe d'égalité* per year

Figure 3.2 further confirms the trend observed in figure 3.1: *principe d'égalité* is increasingly mentioned from 1981 onwards. As mentionned by Pr. Ferey during our exchanges, this coincides with the election of Mitterand, who was politically on the left, while his predecessors were on the right.

3.2.2 Words analysis

Trigrams

CountVectorizer counts the occurrences of words in a document and generates a vector representation using these counts. In contrast, Tfidf considers both the word frequency in the document and its significance in the entire corpus. It assigns greater weights to words that are frequent in a document but rare across the corpus, as they are considered more informative. Tfidf also addresses the influence of common words appearing in multiple documents, which may lack substantial meaning. In summary, CountVectorizer emphasizes word counts, while Tfidf takes into account word importance in both the document and the entire corpus.

In table 3.1 we perform both a CountVectorizer and a Tf-Idf on the utterances that contain *principe d'égalité*. We chose to search for trigrams, as this would enable to identify significant and distinctive trigrams that can effectively represent specific topics or themes associated with *principe d'égalité*.

CountVectorizer		**TfidfVectorizer**	
Term	**Rank**	**Term**	**Rank**
atteinte principe egalite	83	atteinte principe egalite	15.351713
principe egalite suffrage	77	principe egalite suffrage	11.436257
principe egalite devant	53	violation principe egalite	8.873201
contraire principe egalite	43	porte atteinte principe	8.762206
violation principe egalite	42	principe egalite devant	7.833152
porte atteinte principe	34	contraire principe egalite	7.333189
egalite devant loi	30	rupture principe egalite	5.726684
meconnaissance principe egalite	20	meconnaissance principe egalite	5.478087
rupture principe egalite	19	egalite devant loi	5.148882
regard principe egalite	16	objet effet permettre	4.550001

TABLE 3.1: Trigrams for the utterances that contain *principe d'égalité*

This also solidifies the results of the next part, where we analyze the words linked to *principe d'égalité* and their syntactic relations.

3.2.3 POS and dependencies analysis

We performed Part-Of-Speech and dependency analysis to bring out themes associated with *principe d'égalité*, using spaCy's Dependency Matcher (Matthew et al., 2023), in the *fr_core_news_sm* model, on the paragraphs that contain *principe d'égalité*.

We also attempted a comparison with Grew-match [2], a corpus exploration tool for finding and visualizing POS patterns in Universal Dependencies (UD). Unfortunately the corpora do not contain occurences of *principe d'égalité* or even similar use of *principe*.

In our case, and after verification, for every instance of *principe d'égalité* or *principe de l'égalité* we found, *principe* was the head of *égalité*.

[2] Greew-match

Verbs linked to *principe d'égalité*

Head Verbs

In figure 3.3, *Méconnaître* emerges as the most frequent, appearing 23 times. In this context, it signifies the act of miscomprehension or failure to acknowledge the principle of equality.

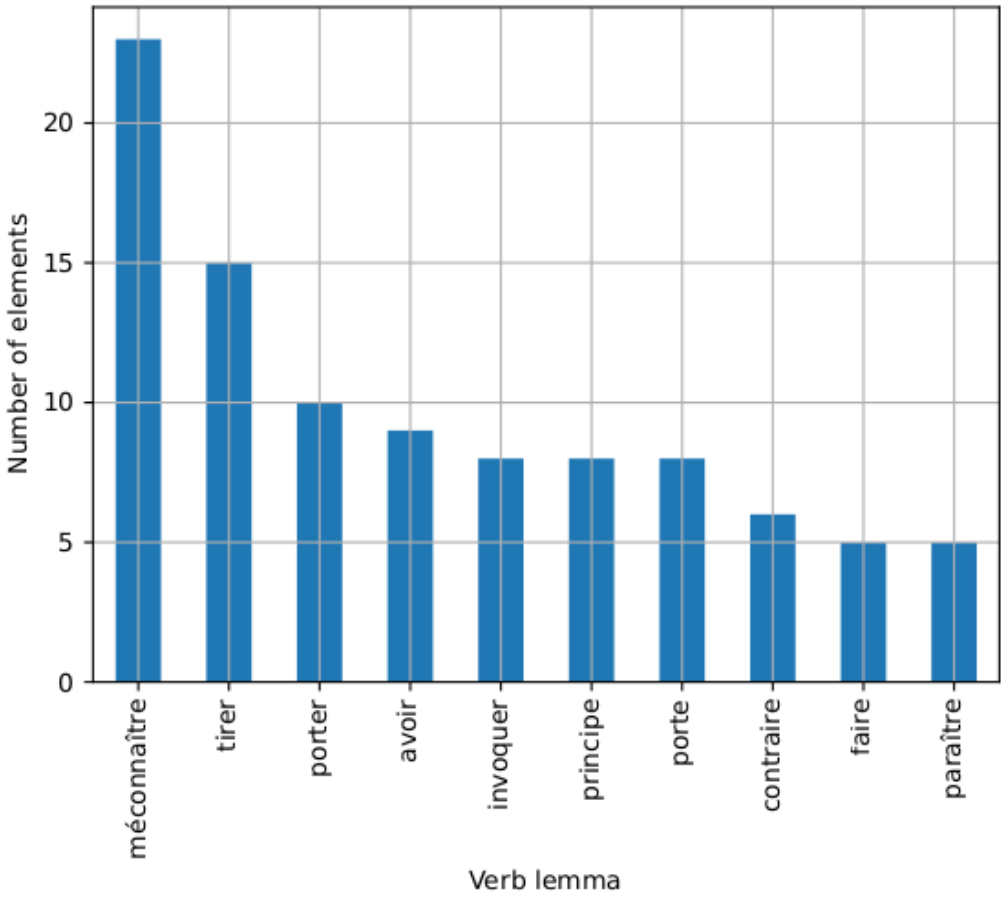

FIGURE 3.3: Distribution of the head verbs of 'principe'

Tirer, with its frequency of 15, embodies the concept of drawing implications or deriving legal consequences from a given set of facts or circumstances.

Porter is used 10 times, reflecting the notion of carrying or bringing forth legal claims or burdens. This value underscores the responsibility entrusted to legal professionals and litigants in presenting their cases and supporting their positions with sound legal arguments and evidence.

The verb *avoir* appears 9 times, reminding us of the essential role of possession and ownership in legal contexts. It encapsulates the notion that legal rights, privileges, and responsibilities are often contingent upon the possession of certain qualities or assets, shaping the dynamics within the legal framework.

With a frequency of 8, *invoquer*, *principe*, and *porte* present themselves as recurring pillars of legal discussions. *Invoquer* emphasizes the act of invoking legal provisions, precedents, or arguments, emphasizing the reliance on established legal frameworks. *principe* highlights the fundamental principles and doctrines that guide

legal reasoning, serving as building blocks for legal decisions. *Porte* evokes the concept of gateways or access points to legal remedies and processes, underscoring the importance of procedural justice.

Contraire emerges 6 times, drawing attention to the presence of contradictions and opposing perspectives within legal debates. It signifies the need to reconcile conflicting viewpoints and find resolution within the legal system, showcasing the intricate nature of legal interpretation and the pursuit of justice.

Finally, *faire* and *paraître* both appear 5 times, offering insights into the presentation and perception of legal arguments. *Faire* relates to the act of undertaking legal actions, while *paraître* points to the way legal arguments or parties may appear to others, emphasizing the role of persuasion and public perception in legal proceedings.

About dependencies

Syntactic dependencies refer to the grammatical relationships between words in a sentence, representing how they are functionally connected to one another. These dependencies are often represented using a dependency tree, where each word is a node and the relationships are depicted as directed edges. Here are examples of some common syntactic dependencies:

- nmod (noun modifier): It represents a noun that modifies another noun. Example: "On examine la constitutionnalité du vote plural, d'une part, au regard de la constitution elle-même, d'autre part, au regard du principe d'égalité devant la loi" - Here, *principe d'égalité* is a noun modifier of *regard*"

- obl:arg (oblique argument): It represents an oblique argument, typically a prepositional phrase, that complements the meaning of the verb. Example: "Il n'est donc pas contraire au principe de l'égalité que la direction d'une banque ou d'une société immobilière soit interdite aux parlementaires." - *principe de l'égalité* is the oblique argument of *contraire*

- obj (object): It represents the direct object of a verb, indicating the entity or thing that is directly affected by the action. Example: "C'est donc le principe de l'égalité des citoyens qui est en cause." - *principe de l'égalité* is the direct object of the verb *être*

- nsubj (nominal subject): It represents the noun phrase that serves as the subject of a clause. Example: "Le principe d'égalité ne permettait, en effet, pas de faire un sort spécial à ces sociétés semblables à celles nationalisées tant par leur statut que par leurs conditions de fonctionnement" - *principe d'égalité* is the nominal subject of the verb *permettre*

- obl:mod (oblique modifier): It represents an oblique modifier, usually a prepositional phrase, that provides additional information about the verb. Example: "Aussi bien, il propose au conseil d'appuyer sa décision, tant sur le principe d'égalité que sur la déclaration des droits de 1'homme de 1789, en tant qu'elle constitutionnalise le droit de propriété." - *principe d'égalité* is the oblique modifier of the verb *propose*

- nsubj:pass (passive nominal subject): It represents the noun phrase that functions as the subject in a passive sentence. Example: "Il suffit, pour que le principe d'égalité soit respecté, que la mesure prise pour le calcul des voix

soit la même à l'égard de tous, chaque fois que le cas envisagé se présente." - *principe d'égalité* is the passive nominal subject of the verb *respecter*

Dependencies linked to principe d'égalité

Table 3.2 demonstrates that, for all dependencies, the top words connected to *principe* are negative and illustrates different aspects of disregarding (*méconnaissance*) or acting in opposition to the principle of equality (*violation*, *contraire*).

Out of the total 411 mentions of *principe d'égalité*, approximately 19% (78 mentions) are associated with *suffrage* as a nmod to *égalité*. This particular nmod construction also accounts for 73.5% (78 out of 106) of all nmods related to *égalité*.

nmod		**obl:arg**		**obj**	
violation	21	contraire	13	méconnaître	12
regard	11	porter	8	mettre	3
méconnaissance	8	porte	7	violer	2
rupture	7	avoir	2	concerne	2
atteinte	5	évident	1	méconnaissent	2
nsubj		**obl:mod**		**nsubj:pass**	
oppose	7	contraire	3	méconnaître	3
faire	4	considérer	1	respecter	1
permettre	2	évident	1	violer	1
contenu	1	propose	1		
être	1	rappeler	1		

TABLE 3.2: Words connected to *principe* and the dependencies by which they are connected

By exploring the children of *égalité*, we also discovered that among the 411 mentions of the *principe d'égalité* there are 78 who have as nmod suffrage (almost 19%), this also represents 73.5% of nmod (78/106) associated with a principle of equality.

In conclusion, we observe the increasing significance of the principle of equality in the Council's debates. Our findings confirm the presence of a distinct lexical field associated with violations of this principle, reflecting the recurring challenges faced by the councillors. Notably, the principle of equality is consistently linked with the concept of voting, as seen in table 3.1, reflecting the Council's frequent engagement with matters related to elections.

These insights shed light on the critical role of the principle of equality in shaping the Council's discussions and decision-making processes. Future research can build upon these findings to further explore the specific implications and applications of this principle within the Council's framework.

Chapter 4

Conclusion

In this project, we focused on three main areas of work: corpus structuring, discussion analysis, and topic modeling centered around the principle of equality. In hindsight, it would have been more beneficial to concentrate our efforts on a specific area, such as corpus structuring as we underestimated the time required to design the corpus structure.

Despite this observation, our contributions in several areas proved to be valuable. Firstly, we assisted in designing a TEI structure that effectively separated different sections of the Constitutional Council's meeting minutes. Additionally, we developed a program capable of identifying speakers, targets, and extracting utterances from the councilors' discussions. This significantly sped up the document processing phase. We also presented an initial approach for annotating interactions among the advisors, including their agreements and disagreements. Our method for predicting these interactions shows promise within its intended context. Lastly, we analyzed instances of the "principle of equality" concept in order to gain insights into its evolving impact on the councilor's debates.

This project provided us with valuable lessons regarding the acquisition of complex corpora. We acquired knowledge of the TEI format and its associated challenges, and we gained hands-on experience with an NLI model, particularly in hypothesis design considerations. Moving forward, we recognize the importance of prioritizing specific areas of focus to ensure more in-depth and high-quality research outcomes. With the lessons learned from this project, we are well-equipped for future endeavors in this field.

Appendix A

Agreement visualization

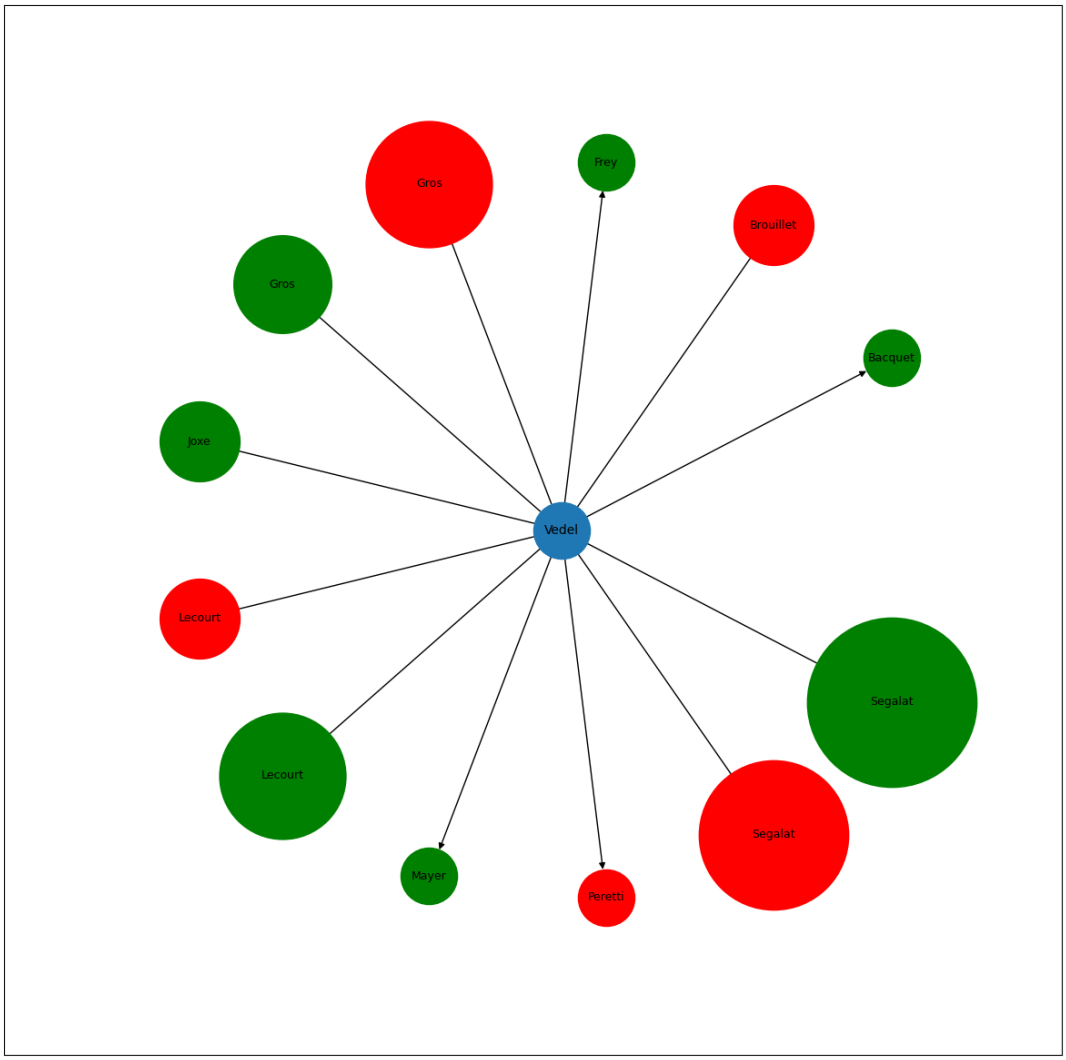

FIGURE A.1: Agreement occurences graph with Mr. Vedel as speaker. Green represent agreement while red represent disagreement.

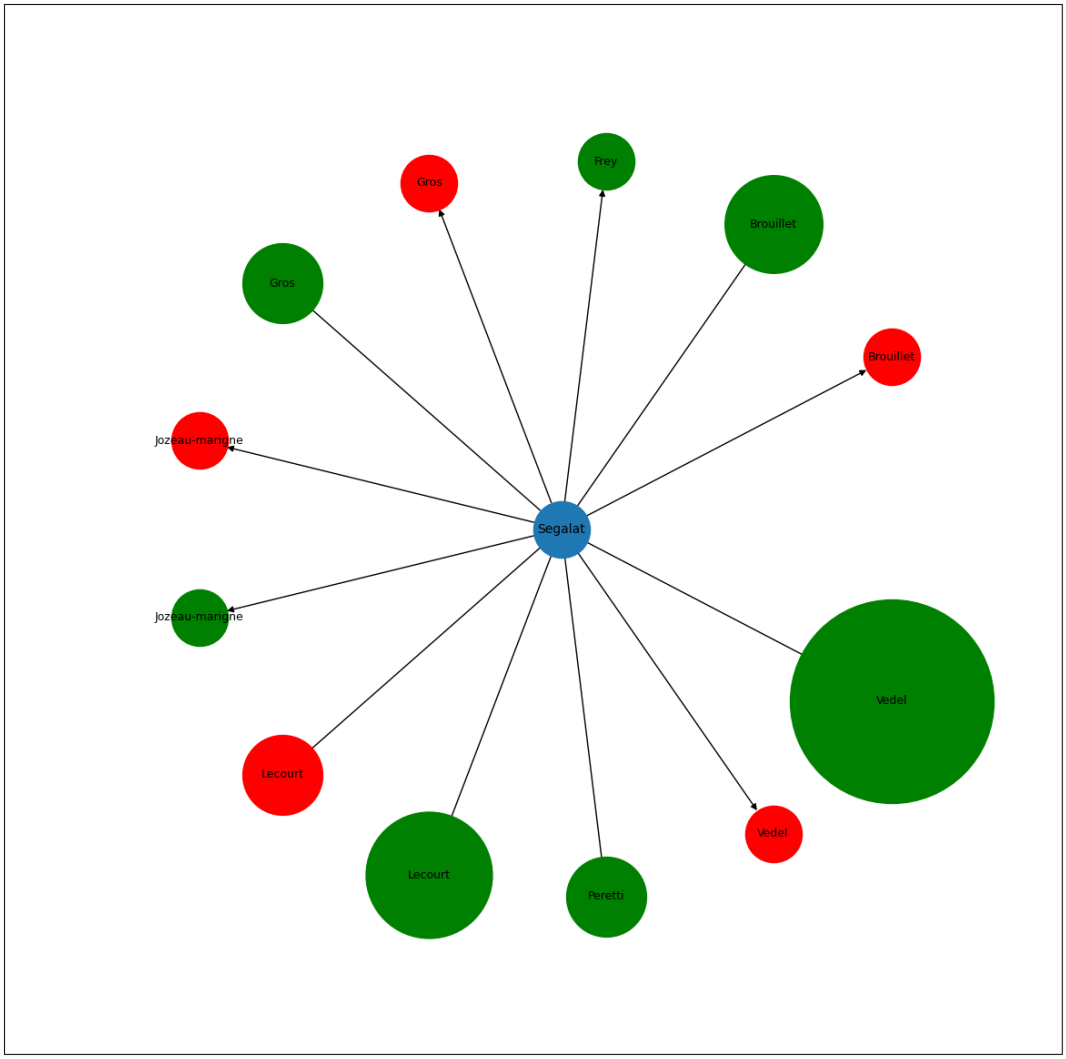

FIGURE A.2: Agreement occurences graph with Mr. Segalat as speaker. Green represent agreement while red represent disagreement.

Bibliography

Conneau, Alexis et al. (2018). "XNLI: Evaluating Cross-lingual Sentence Representations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485. DOI: 10.18653/v1/D18-1269. URL: https://aclanthology.org/D18-1269.

Erjavec, Tomaž and Andrej Pančur (Sept. 2019). *Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings*. DOI: 10.5281/zenodo.3446164. URL: https://doi.org/10.5281/zenodo.3446164.

Grootendorst, Maarten (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv: 2203.05794 [cs.CL].

MacCartney, Bill and Christopher D. Manning (Aug. 2008). "Modeling Semantic Containment and Exclusion in Natural Language Inference". In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, pp. 521–528. URL: https://aclanthology.org/C08-1066.

Martin, Louis et al. (July 2020). "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645. URL: https://aclanthology.org/2020.acl-main.645.

Matthew, Honnibal et al. (May 2023). *spaCy: Industrial-strength Natural Language Processing in Python*. DOI: 10.5281/zenodo.7970450. URL: https://doi.org/10.5281/zenodo.7970450.

Williams, Adina, Nikita Nangia, and Samuel Bowman (June 2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: https://aclanthology.org/N18-1101.

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.