HAL open science

# Querying a large annotated corpus of parliamentary debates

Sascha Diwersy, Giancarlo Luxardo

## To cite this version:

Sascha Diwersy, Giancarlo Luxardo. Querying a large annotated corpus of parliamentary debates. LREC, ParlaCLARIN Workshop, 2020, Marseille, France. hal-03317717

**HAL Id: hal-03317717**

**https://hal.science/hal-03317717v1**

Submitted on 7 Aug 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

*Proceedings of ParlaCLARIN II Workshop*, pages 75–79
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

# Querying a Large Annotated Corpus of Parliamentary Debates

**Sascha Diwersy, Giancarlo Luxardo**
Praxiling UMR 5267 (Univ Paul Valéry Montpellier 3, CNRS)
sascha.diwersy@univ-montp3.fr, giancarlo.luxardo@univ-montp3.fr

**Abstract**
The TAPS corpus makes it possible to share a large volume of French parliamentary data. The TEI-compliant approach behind its design choices facilitates the publishing and the interoperability of data, but also the implementation of exploratory data analysis techniques in order to process institutional or political discourse. We demonstrate its application to the debates occurred in the context of a specific legislative process, which generated a strong opposition.

**Keywords:** political discourse, parliamentary corpora, metadata, cooccurrences.

## 1. Introduction

The present paper describes the current version of a family of parliamentary corpora called *Transcription and Annotation of Parliamentary Speech* (TAPS). A previous publication introduced the methodology adopted to set up the corpora and the basic software components (Diwersy *et al*., 2018). After a reminder of the corpus structure and of the technologies implemented, we focus on an example of application, the debates about a proposed legislation, involving both text retrieval functionalities and the processing of the extracted lexicons through two data analysis techniques: correspondence analysis and specificity analysis.

## 2. Text segmentation and linguistic annotation

TAPS-fr is a corpus (or more precisely a family of corpora) allowing the access to the complete transcription of the French parliamentary debates in plenary sitting. It is compiled making use of the XML-based open data published on the Web site of the *Assemblée Nationale*, transformed through multiple steps.

The TAPS format is a compromise stemming from the use of several formats:

- the metadata extracted from the source open data (based on an undocumented model),
- the TEI guidelines for the transcription of oral corpora,
- the components of the IMS Open Corpus Workbench (CWB), and in particular the Corpus Query Processor (CQP) (cf. Evert & Hardie, 2011).

A CWB corpus is stored according to a tabular format (token-based), which encapsulates an XML mark-up. CQP tools allow to produce frequency counts coded within data tables. These tables can be processed using statistical procedures (usually in the R environment). The integration of common open source software (CWB, R) facilitates the data interchange and the experimentation with various tools. Among other tools integrating the aforementioned technologies, the TXM textometry software, developed in the French communities of digital humanities and discourse analysis, was used to process the TAPS-fr corpus. In addition, TXM provides functionalities for a Web-based publication (TEI-compliant) of the corpus, which makes possible the online access of the TAPS-fr corpus[1].

The XML encoding used for TAPS is basically the one described by the TEI <u> (utterance) element included in the module: *Transcription of speech*. In our context, the segment applied in the scope of <u> is not a single utterance, but the text portion determined by the change of speaker (the speaker's turn). A number of attributes are added to the <u> element, describing the speaker (name, party, role in the debate, etc…). The repetition of this metadata for every single speech leads to some redundancy but allows a fast text retrieval. The identification of the sitting and its date are instead found at the top level of the tree or in the TEI header (multiple sittings may occur in the same day). TEI also allows to describe paralinguistic events (incidents associated to a speech, such as noise or interruptions).

The CQP environment distinguishes two annotation levels related to the units generated by the compilation process:

- the structural units are those provided by the text tokenization (and in our case derived from the TEI encoding), they describe both the text semantics and some formatting characteristics,
- the lexical units represent the linguistic annotation and are added to each token in the text.

Two optional annotation modes have been experimented with TAPS for the linguistic annotation:

- morphosyntactic tagging and lemmatization by means of TreeTagger (lemma + part-of-speech), *cf.* (Schmid, 1994),
- syntactic analysis, with additional features, in particular related to dependency relations (the Bonsai pipeline was experimented), *cf.* (Candito *et al*. 2010a, 2010b).

In the remainder of this paper, we concentrate on the description of procedures, which can be basically implemented within the CWB and R environments, independently of higher-level tools. The TreeTagger option was chosen.

## 3. Use scenario

We now demonstrate some analysis techniques that can be performed on TAPS, taking as an example the debates held

[1] https://textometrie.univ-montp3.fr/

in the course of the review of the law named « loi Travail » or « loi El Khomri », adopted on the 8th of August 2016. The presentation of this law, which aimed to simplify the French Labor Code in order to reduce unemployment, provoked numerous popular protests in the country and the discussions in parliament, started in February 2016, entailed a split in the majority previously supporting the government. These protests, including spontaneous demonstrations (such as those known as *Nuit debout*) or strikes initiated by trade unions or student organizations, were supported by part of the left: left parties not participating in the government, but also by members (also known as *frondeurs*) of the Socialist Party, the leading majority party, as well as some ecologists (while their party, *Europe Écologie – Les Verts*, was still in the majority).

It is then interesting to question the vocabularies used by the different political parties during the related sittings. More specifically, questions relevant for a political analysis are about the cohesion of the discourses within the majority parties (socialists, radicals, ecologists) represented in the government and the possibility to find out unexpected proximities with other parties.

## 4. Collocational analysis based on TAPS-fr-2

The analysis is performed against the corpus named TAPS-fr-2, covering the period April 2012 – February 2017 and totaling about 28 million occurrences of tokens. Although functionalities are available to extract a subcorpus of a smaller size (faster to process), the search is here performed over the full corpus: the resulting frequency distributions then include the occurrences of other periods of debates.
The following approach is taken:

- starting from a CQP query, a lexical table comprised of the collocates (represented as lemmas) appearing within the same utterance in the left and right co-text of the node "loi travail" or "loi El Khomry" (including possible variants)[2] is built: the list is restricted primarily to nouns, proper nouns, adjectives, adverbs and verbs;
- various thresholds for the minimum co-frequency of the collocates are tested;
- various statistical tests are applied to measure the significance level of each collocate with each political group (Fisher's exact test being the first choice);
- another table "cross-tabulating" these collocates and the political groups of the speakers related to the utterances is generated (seven political groups are identified, including non-attached members), filtering the rows according to either the test score or the minimum threshold;
- a correspondence analysis is performed on the cross-tabulation;
- for each party, the most characteristic collocates are provided by means of a specificity analysis.

## 5. Results

The following political groups are considered:
- Écolo: Groupe écologiste
- GDR: Gauche démocrate et républicaine
- NI : Non inscrits
- RRDP : Radical, républicain, démocrate et progressiste
- SRC_SER : Socialiste, républicain et citoyen / Socialiste, écologistes et républicain
- UDI : Union des démocrates et indépendants
- UMP_LR : Union pour un Mouvement Populaire / Les Républicains [3]

The table cross-tabulating groups and lemmas is produced using a minimum threshold of 10 for the co-frequency count. It contains 5313 rows.

The application of a correspondence analysis (CA) produces results displayed by Figure-1[4], which illustrates the distance between the parties.

According to the contributions generated by the CA, the first axis shows an opposition mainly between SRC_SER (socialists) and Écolo (ecologists). On the second axis, the opposition is between GDR and Écolo.

The graphic represented by Figure-2 also shows the most contributive lemmas.

The most characteristic lemmas of each group can be highlighted by the study of the results of the CA, but also by means of the computation of frequency specificities[5]. This technique, based on the hypergeometric distribution, is described by (Lafon, 1980). The three bar plots displayed by Figure-3 show the contrasts between the various parties of the TAPS corpus based on the 10 most characteristic lemmas of the three mostly contributing groups to the CA.

The interpretation of the presence of the items associated to each political group requires some caution. Verifications of the related co-text (e.g. with a concordance function) are necessary, sometimes revealing collocations that result from recurrent formulaic expressions. The following comments may be attempted:

1. The socialists (SRC_SER) focus on several details of the content of the law (*formation, compétence, assurance, permis*[6]) as well as the legislative procedure (*amendement, validation*[7]).

[2] The CQP query expression we used to identify the node is as follows (with frlemma and frpos representing the (positional) attributes lemma and PoS): [frlemma="loi" %cd] [frpos="PUN.*" %cd]? [frlemma="travail" %cd] | [frlemma="loi" %cd] []? [frlemma="relatif|relative|sur" %cd] [] [frlemma="travail" %cd] | [frlemma="loi" %cd] []? [word="el" %cd] []? [word="k.*o.*" %cd]

[3] SRC and UMP groups have changed their name during the legislature.

[4] We used the R packages *FactoMineR* (Lê et al., 2008) to compute the CA, and *explor* (Barnier, 2017) to generate the plots shown in Figures 1 and 2.

[5] We computed the specificity scores by means of the R package *textometry* (Heiden, 2010).

[6] English: training, competency, insurance (the English *assurance* would be unlikely here), permit

[7] as in English

2. The discourse of the ecologists appears related to the *Nuit debout* movement, in connection with the adoption of the law: *mouvement, contestation, précarité, démocratie, manifester, mobilisation*[8]. The presence of the word *urgence* suggests an interference with the situation of state of emergency (*état d'urgence*) declared in France after the attacks of November 2015.
3. The characteristic items of the left-wing opposition (GDR) are instead related to the social aspect of the law and its expected consequences: *temps, partiel, pauvreté, salaire*[9]. Other words are related to arguments assuming a relationship with the European treaties: *plan, traité*.

The position of the Écolo group opposed to SRC_SER is an unexpected result, which deserves a thorough examination of the related contexts. It appears that the retrieved speakers' turns for the ecologists are only seven, related to four members of the parliament (in four different sittings), all of them critical of the law or the process to adopt it. It must also be noted that the relatively small volume of contributions associated to the ecologists is explained by the fact that the group was dissolved in May 2016, as six members decided to join the socialist group.

The results of our analysis do not highlight a specific critical trend within the socialists with respect to the government, which can be explained by the smaller number of contributions of the most critical members of the group (four of them leaving the group during the legislature). However, more detailed observations at the level of individual members of the parliament should be performed.

## 6. Conclusion

In this paper, we have described an application scenario of the TAPS-fr corpus involving a large volume of parliamentary data. We consider that within the analytical framework of textometry and with the common tools of the corpus linguistics area it is possible to make an effective use of the resource. Similar approaches can be adopted on different use scenarios, including those based on other variables (e.g. time or speaker status), in addition to party affiliation. While the presented scenario has the advantage to make use of a meaningful participation, in terms of volume of data and contrasted positions, it would be interesting to consider other types of scenarios, more technical and possibly more challenging for the methods here demonstrated.

While the resource is already published and openly accessible, efforts need to be undertaken in order to improve its dissemination and long-term preservation. Future developments also include new features allowing a continuous expansion of the corpus, with the latest sessions of the assembly, and possibly as well extensions to additional institutions.

## 7. Bibliographical References


Barnier, J. (2017). *explor: Interactive Interfaces for Results Exploration* (Version 0.3.3). Retrieved from https://CRAN.R-project.org/package=explor

Diwersy, S., Frontini, F., Luxardo, G. (2018). The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse, in *Proceedings of ParlaCLARIN workshop, 11th edition of the Language Resources and Evaluation Conference (LREC2018)*.

Candito, M., Crabbé, B., and Denis, P. (2010a). Statistical French dependency parsing: treebank conversion and first results. In *Seventh International Conference on Language Resources and Evaluation - LREC 2010*, pages 1840–1847, La Valletta, Malta, May. European Language Resources Association (ELRA).

Candito, M., Nivre, J., Denis, P., and Henestroza Anguiano, E. (2010b). Benchmarking of Statistical Dependency Parsers for French. In *23rd International Conference on Computational Linguistics - COLING 2010*, pages 108–116, Beijing, China, August. Coling 2010 Organizing Committee.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, Birmingham, UK.

Evert, S. (2019). *CQP Query Language Tutorial*, CWB Version 3.4.16. http://cwb.sourceforge.net/

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In K. I. Ryo Otoguro (Ed.): *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24* (p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Lavrentiev, A., Heiden, S., Decorde, M. (2013). Analyzing TEI encoded texts with the TXM platform. *The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013*, Oct 2013, Rome, Italy. halshs-01118120

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1–18.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.


[8] English: movement, contestation, precarity, democracy, protest, rallying

[9] part, time, poverty, salary

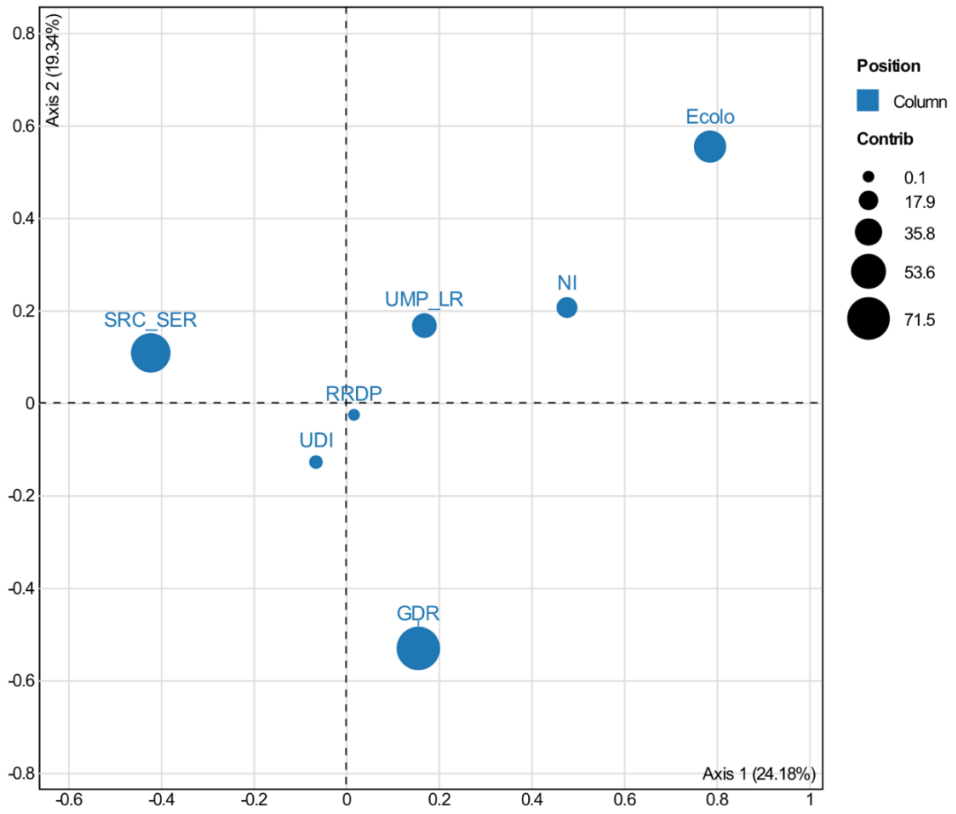



*Figure 2- CA (rows hidden)*

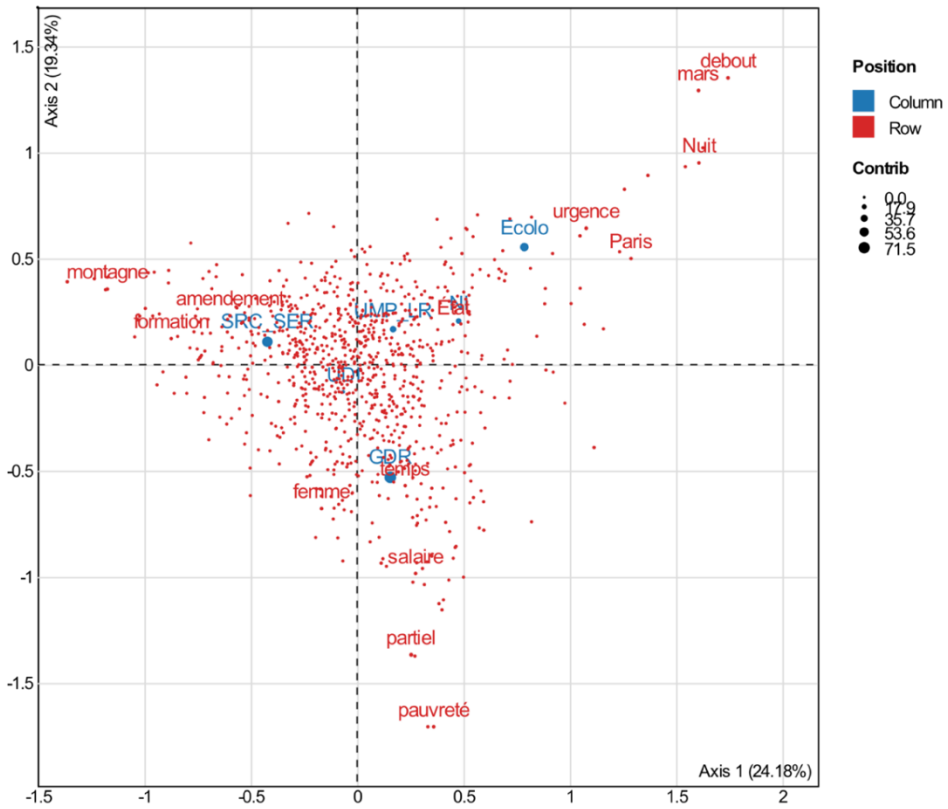
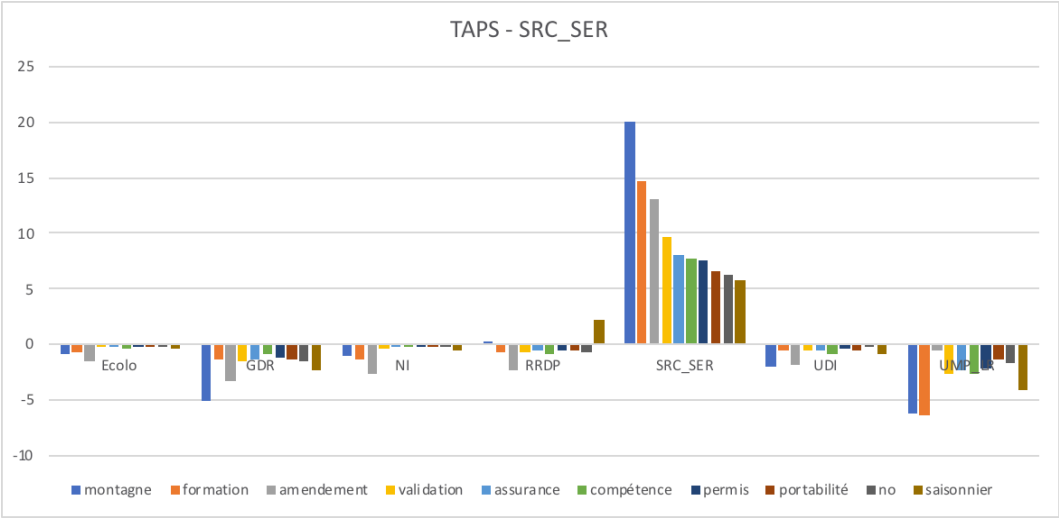
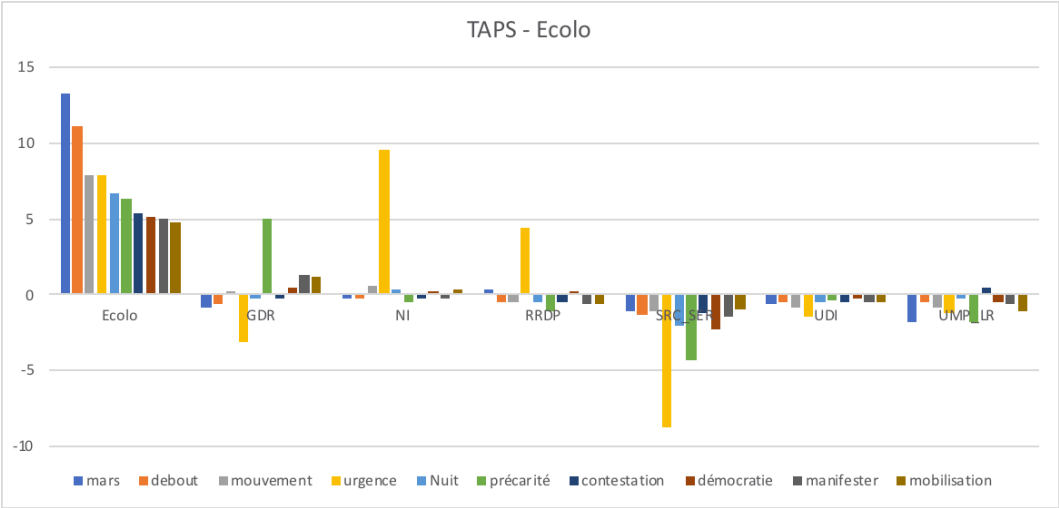



*Figure 1 - CA (displaying most contributive rows)*



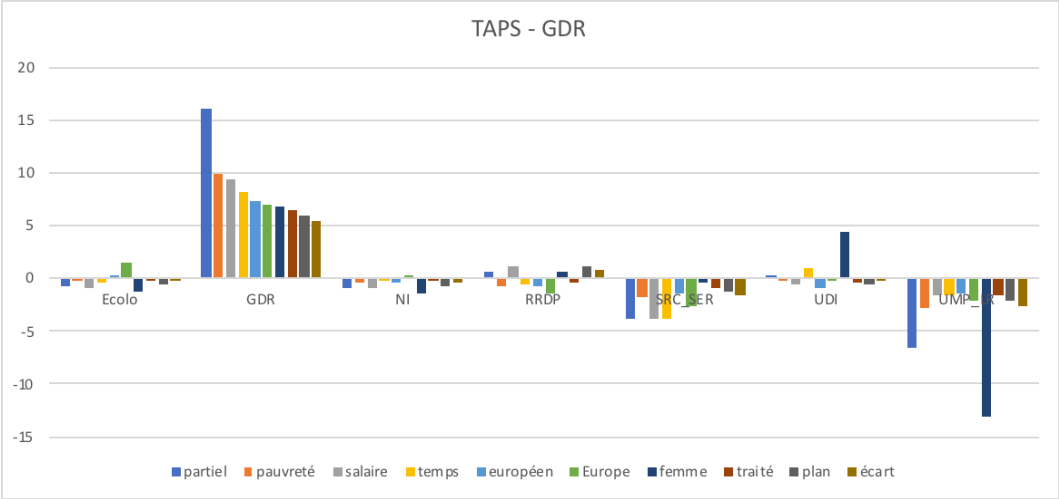




*Figure 3 - Specificity analysis for three groups*