The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse

Sascha Diwersy, Francesca Frontini, Giancarlo Luxardo
PRAXILING UMR 5267 Univ Paul Valéry Montpellier 3 & CNRS - Montpellier, France
name.surname@univ.montp3.fr

Abstract

This paper reports on a corpus collecting together the French parliamentary debates in plenary sitting. It outlines the design and data format of the samples and presents various usage scenarios related to their textometric use.

Keywords: political discourse, parliamentary corpora, metadata

1. Introduction

This contribution presents a corpus that contains the transcriptions of French parliamentary debates; we also discuss the possibilities of its exploration from a textometric perspective.

The paper is structured as follows. Section 2 outlines the rationale behind the creation of the corpus, its composition and the concepts involved in its design. Section 3 describes how the corpus can be processed to be implemented in a publishing platform. In section 4, we then introduce some key elements of a textometric methodology and illustrate the exploration of the corpus by giving brief sketches of corresponding usage scenarios. Section 5 provides a summary and discusses possible directions for future developments of the resource presented in this paper.

2. The TAPS-fr corpus

2.1. Rationale and composition of TAPS-fr

The *Assemblée Nationale* publishes on an open access basis[1] a number of datasets, and among them its debates in plenary sitting, dating back to 2013 and provided as a regularly updated data dump at `http://data.assemblee-nationale.fr/travaux-parlementaires/debats`.[2].

The textual data have been processed and assembled according to a methodology discussed in the following and called *Transcription and Annotation of Parliamentary Speech* (TAPS), with the aim of offering a large-scale resource to researchers working on French political discourse, especially from a data-driven perspective. The methodology is kept as generic as possible, in order to be reused for debates of additional parliaments, possibly in other languages.

We call the corpus described here TAPS-fr. It is primarily designed to provide a methodological support for investigations in the French tradition of textometry (French: *textométrie*), which integrates both searches based on full-text indexing and multivariate exploratory data analysis (Lebart et al., 1998). The open data publishing of the French parliamentary debates is part of a trend known as Open Government Data and described with the eight principles defined by the Sebastopol meeting held in 2007[3]. One of the challenges for the projects initiated in this trend is set by the fact that these open data, while published in large amounts and accessible with a relative ease of reuse, are "raw data": little is known about the conditions of their production (Plancq, 2016).

Legislature	Period	Nr of sessions	Nr of words
14	05/13-12/13	152	5,200 K
14	01/14-02/17	873	28,600 K
15	06/17-12/17	156	4,700 K
Total			38,500 K

Table 1: Composition of the TAPS-fr

We subdivided the TAPS-fr corpus into three subcorpora described by Table 1:

1. The first months (May 2013 - December 2013) represent a small subcorpus, which was not processed in depth so far (the source webpage states that the debates were fully transcribed only from October 2013).
2. The second subcorpus was the one mostly used for our experiments: it comprises the debates of the last months of the 14th legislature (January 2014 - February 2017).
3. A third corpus includes the debates of the 15th legislature up to the end of December 2017.

2.2. A machine-processable format geared to multiple needs

We distinguish four formats handled for the processing of TAPS-fr, all of them being XML-based:

1. The source format: it is the format used by the raw data, which is subdivided in three components (actors, bodies - *organes* - and sittings); the text is included in the sittings section and refers to actors (members of

[1] The "Licence ouverte / Open Licence" is a free licence created by the French governmental mission Etalab.

[2] A selection of parliamentary records from the *Assemblée Nationale* has already been collected and published in TEI format in (Truan, 2017) as part of a broader project on perceptions of the other in various European countries.

[3] https://opengovdata.org/

parliament, members of government, etc.) belonging to various bodies (e.g. parliamentary commissions).

2. The TAPS format: it is the result of a conversion applied to the source data, in order to extract the relevant metadata and annotation useful to our applications (see below).
3. The XML-TXM format is a customization of the TEI data model used by the TXM software as a pivot format in order to present semantic and editorial annotations.
4. The CWB format (defined by the IMS Open Corpus Workbench) encapsulates lexical and syntactic annotations and is used by a search engine, allowing text retrieval based on queries expressed in the CQP (Corpus Query Processor) syntax. This is a compound format with XML tags and token records appearing on separate lines (one surface form is associated to tab-delimited token-level annotations).

The descriptions of the TXM (for "textometry") platform (Lavrentiev et al., 2013) and of the CWB environment (Evert and Hardie, 2011) are beyond the scope of the present document. However, the following two sections describe how they have been implemented in this project.

The TAPS format basically relies on the concepts of metadata and annotations defined in the TXM environment, which distinguishes structural units and lexical units.

Metadata (the association of a variable and a set of modalities) are used to partition the corpus, to create subcorpora and to retrieve the text. They are defined on various structural units: XML elements, text segments, paragraphs, sentences, and possibly other units defined by the user. Metadata are therefore viewed as properties of structural units.

Each processed lexical unit has several properties, such as word form, lemma and part-of-speech (grammatical category).

The conversion from the source format to the TAPS format creates a number of files, each one associated to a parliamentary sitting. The metadata that is extracted from the source format can be associated either globally to each file or to a single speech (the intervention of a person in the debate). The format of the files complies with the TEI data model, so that the metadata associated to a file are described in the TEI header. For the single speech, the `<u>` element (utterance) was chosen[4]: this element is originally defined by the TEI guidelines for the transcription of oral corpora, it is extended by the definition of a number of attributes relevant to our application (e.g.: role in the debate, nomination in the parliamentary structures, nomination in the government, political affiliation...). The assignment of attributes at the utterance level and within the TEI headers implies some redundancy, however it provides an easier reuse for the text retrieval. Table 2 specifies the major structural units defined within the data model of the TAPS-fr corpus.

[4]This approach was also adopted by (Truan, 2017) and by the authors of the SlovParl corpus (Pančur et al., 2017), whereas earlier versions of the latter opted for the `<sp>` element defined by the TEI module for encoding performance texts (cf. `https://github.com/SIstory/SlovParl`).

Structural Unit	Associated Metadata (descriptors)	XML Element
sitting	date-time, year, parliamentary term	`<text>`
speech	speaker name, speaker role, parliamentary group, speech type (debate, interruption, vote explanation, etc.)[5]	`<u>` (utterance)
paralinguistic event	description	`<incident>`
sentence[6]	–	`<s>`

Table 2: Main structural units encoded in the TAPS-fr corpus

Starting from the TAPS format, the TXM environment performs several steps of conversion and generates files in the XML-TXM format as well as in the CWB format, which, in both cases, can include linguistic annotations added to the lexical units. While TXM's import modules allow for automatic morphosyntactic tagging and lemmatisation by means of TreeTagger[7] (Schmid, 1994), it is possible to pre-process the corpus data outside the platform by using other NLP toolkits. In our specific case, we chose the freely available processing pipeline Bonsai[8] (Candito et al., 2010b; Candito et al., 2010a) in order to add syntactic dependency annotations. The latter have been extended by several categories whose purpose is to optimize the processing of queries exploring the dependency relations annotated in the corpus. The categories in question are marked by an asterisk in Table 3, which outlines the overall data model of the word level annotations within the TAPS-fr corpus.

3. The publishing framework

The four formats described in section 2.2 enable to publish the TAPS-fr corpus in two different contexts: either within the TXM interface or in the TAPS format for the purpose of improving interoperability. Both options provide some conformance level with the TEI guidelines.

Unlike older software tools developed within the French community of "*analyse des données textuelles*" (textual data analysis), TXM was designed to support applications

[5]From a linguistic point of view, this descriptor, which is not included in the data model of the corpus provided by (Truan, 2017), is particularly important when it comes to differentiate effects of register variation ranging from highly formulaic to less formal speech (as in the case of e.g. interruptions). It should be noted that this metadata element can be easily retrieved from the raw data dump we used to build the TAPS-fr corpus, whereas it is only partially or not all available in the other source formats (HTML or pdf) used by the *Assemblée Nationale* to publish its minutes on-line.

[6]This unit is optional as it is only provided in the case of specific processing steps pertaining to linguistic annotation.

[7]`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[8]`https://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`

Lexical Property	Description
word	surface form or punctuation sign
lemma	lemma corresponding to the surface form
cpos	coarse grained part of speech (PoS)
pos	fine grained PoS (+ subcategorization)
feat	morphological features
deprel	syntactic function of the token in the dependency relation to its head
headword *	surface form of the syntactic head
headlemma *	lemma of the syntactic head
headcpos *	coarse grained PoS of the syntactic head
headpos *	fine grained PoS (+ subcategorization) of the syntactic head
headfeat *	morphological features of the syntactic head

Table 3: Linguistic annotations at the word level

for textual criticism. The ability to produce a critical edition of a historical source, typically a "synoptic edition". i.e. a formatted output presented alongside a facsimile (e.g. a manuscript), is provided by the use of the TEI guidelines.

The TXM conformance to TEI is implemented through the use of the TXM pivot format (XML-TXM[9]), which is basically derived from the need to generate a document to be rendered in a browser (with specific layout directives), while allowing the navigation in the text through the support of the TEI <w> element (which encapsulates morphosyntactic and other annotations of the words).

In the context of TAPS-fr, as a parliamentary corpus, the basic requirement is to provide a single edition allowing navigation in a browser and keeping the editorial annotations made by the transcribers. Possible extensions would be to provide multiple editions including: the translation to a different language (but this usage would be unlikely in the case of the *Assemblée Nationale*) or links to audio or video recording.

The TAPS-fr corpus is available from the textometry portal of the Praxiling laboratory [10], either directly browsable online in the TXM environment or as a downloadable resource (with the possibility to process it offline in the desktop version of the TXM software).

In order to comply with the TEI model, the TAPS sitting files (already mentioned above) contain a `TEIheader`: apart from the information related to the publication conditions (`<fileDesc>`), this header also describes the date of the sitting and the speakers involved (`<creation>` element in `<profileDesc>`).

[9]In addition to the generic XML format, TXM also integrates TEI with an import module, called TXM-XTZ (XML TEI Zero), which is able to interpret the semantics associated to a minimal set of TEI elements, through the application of XSL stylesheets.

[10]http://textometrie.univ-montp3.fr/

4. The analytical framework

In this section we will briefly illustrate the application of two standard methods in textometry - correspondence analysis (CA) and the identification of characteristic items by frequency specificities - to the TAPS-fr corpus. Correspondence analysis (cf. (Benzécri, 1973), (Lebart et al., 1998, 45ssq)) is a useful technique in providing a condensed view of divergences relating to samples (resulting from a partition in the corpus) and lexical items.

We illustrate this by means of a plot generated on the basis of a CA (Figure 1)[11] performed on the speeches in the second subcorpus (cf. section 2) using the political group of each speaker (excluding the sitting presidents and the members of government) as differentiating variable. It is possible to observe that the first (horizontal) axis opposes the right-wing groups (UMP-LR, UDI), which have negative coordinates, to the left-wing groups (SRC-SER, Écolo, RRDP, GDR), which are located on the positive side, whereas on the second (vertical) axis, the socialist group (SRC-SER), which forms the major part of the government majority during that period, stands in contrast to the group of left-wing opposition parties GDR.

An efficient way to single out the lexical (and grammatical) items implicated in the opposition of extralinguistic factors highlighted by CA is the computation of frequency specificities based on the hypergeometric distribution (Lafon, 1980), a lexico-statistical approach similar to the keyword analysis used in the British tradition of corpus linguistics (cf. amongst others (Rayson, 2003)). Figure 2 highlights some of the nouns that are more specific to the discourse of the right-wing parliamentary group *UMP-LR*. These include nouns referring to the nation (*Français* 'French') and other classic elements of conservative ideology both in social (e.g. the series *famille* 'family', *parent* 'parent', *enfant* 'child') and economic terms (e.g. nouns designating learned professions such as *médecin* 'doctor, physician', *notaire* 'notary, solicitor').

5. Conclusion and outlook

In this paper, we introduced TAPS-fr, a corpus of debates from the *Assemblée Nationale* by giving a brief sketch of the methodology underlying its creation, of its data model and of some application scenarios that illustrate the exploitation of this resource within the analytical framework of textomtry. The corpus, whose preliminary version is now accessible at `http://textometrie.univ-montp3.fr/`, will be published in its stable version in the Ortolang[12] CLARIN repository for long term preservation, under the CC BY-NC 4.0 license. As the TAPS-fr is meant to be a monitor corpus[13], it will continually be expanded on the basis of the regularly updated raw

[11]The CA plot was generated by means of the R packages FactoMineR (Husson et al., 2013) and explor (Barnier, 2017). We have chosen these packages instead of TXM's CA command because they allow for a more flexible manipulation of the graphical output. The axis descriptions indicated by the horizontal and vertical arrows have been added in a post-processing step.

[12]`https://www.ortolang.fr/`

[13]For the notion of monitor corpus see amongst others (McEnery and Hardie, 2011, 6sq.).

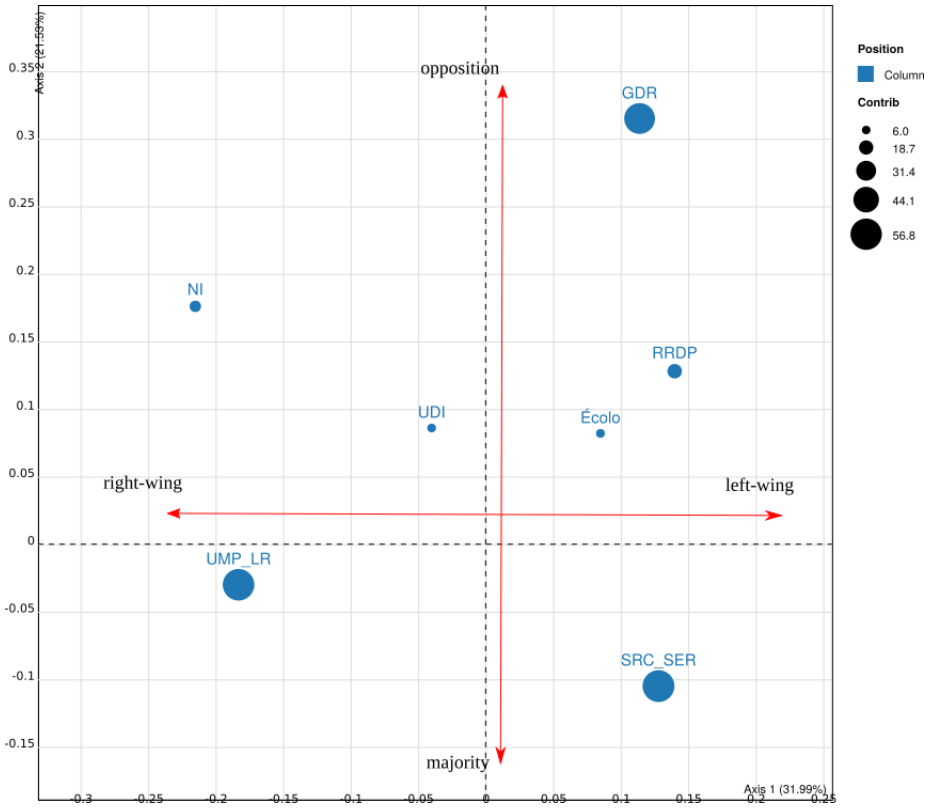

Figure 1: CA plot based on a partition by political group (word-rows are not displayed). Point sizes indicate contribution.

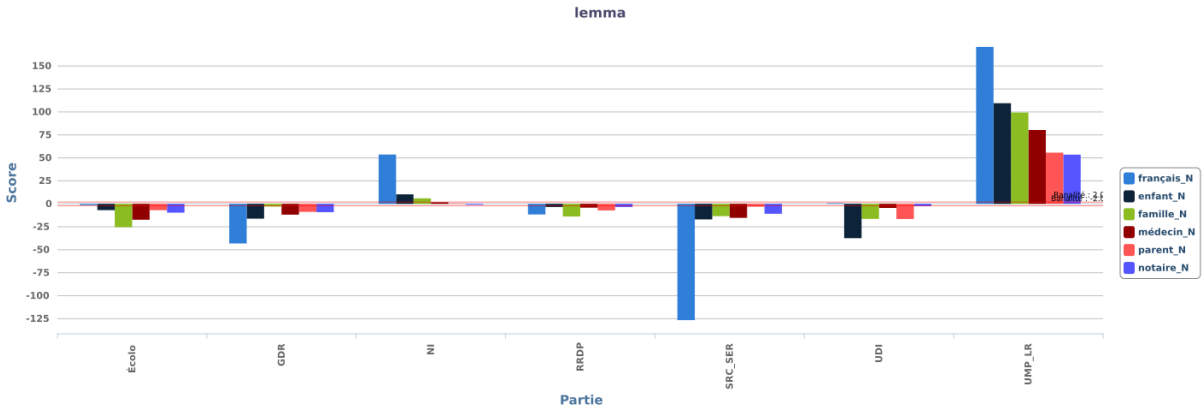

Figure 2: Most characteristic nouns specific to the discourse of the parliamentary group UMP-LR

dump provided by the *Assemblée Nationale*'s open data service, but we also intend to include successively samples of parliamentary sessions prior to 2013. In the latter case, this implies the necessity to adapt our current approach to the processing of material coming from various source formats (primarily HTML and XML with different schema specifications) with varying granularity of directly retrievable information, which might lead to slight revisions of the data model presented in this paper. The future stabilization of our methodology could lay ground not only to the continuous construction of a resource providing broad coverage of the debates at the French *Assemblée Nationale*, but also to the project of creating an extended textual base, which by integrating the plenary sessions of the *Sénat*, the second chamber of the French Parliament, would constitute a large scale corpus of institutional and political discourse in contemporary France at the national level.

6. Bibliographical References

Barnier, J., (2017). *explor: Interactive Interfaces for Results Exploration*. R package version 0.3.3.

Benzécri, J.-P. et collaborateurs. (1973). *L'analyse des données : L'analyse des correspondances*, volume 2. Dunod Paris.

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.

Husson, F., Josse, J., Lê, S., and Mazet, J. (2013). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. Technical report.

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., and Decorde, M. (2013). Analyzing TEI encoded texts with the TXM platform. October.

Lebart, L., Salem, A., and Berry, L. (1998). *Exploring Textual Data*, volume 4 of *Text, Speech and Language Technology*. Springer Netherlands, Dordrecht.

McEnery, T. and Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Pančur, A.,
v Sorn, M., and Erjavec, T. (2017). Slovenian parliamentary corpus SlovParl 2.0. Slovenian language resource repository CLARIN.SI.

Plancq, C. (2016). Utiliser les données ouvertes (open data). Un exemple avec les débats en séance publique à l'Assemblée nationale.

Rayson, P. (2003). *Matrix:a statistical method and software tool for linguistic analysis through corpus comparison*. phd, Lancaster University, February.

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

Truan, N. (2017). Débats parlementaires sur l'europe à l'assemblée nationale (2002-2012). ORTOLANG (Open Resources and TOols for LANGuage) –`https://hdl.handle.net/11403/fr-parl/v1`.