CLARIN

ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora

Miyazaki, 7 May 2018

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

Parliamentary debates in France

The ***Assemblée Nationale*** publishes on an open access basis a number of datasets, and among them its debates in plenary sitting, dating back to 2013.

http://data.assemblee-nationale.fr/travaux-parlementaires/debats

The TAPS-fr corpus

From this source data the TAPS-fr (**Transcription and Annotation of Parliamentary Speech**) corpus was derived.

Keep the methodology as generic as possible, in order for it to be reused for debates of additional parliaments, possibly in other languages.

Content of TAPS-fr

Législature (term)	Period	Nr of sittings	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**

Composition

- 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).
- **The second subcorpus was the one mostly used for our experiments: it comprises the debates of the last months of the 14th “législature” (January 2014 - February 2017).**
- A third corpus includes the debates of the 15th legislature up to the end of December 2017.

The formats

- The source format - subdivided in three components (actors, bodies - *organes* - and sittings)
- TEI-XML format
  - import into **TXM** (open-source text/corpus analysis environment)
- CWB format - **IMS Open Corpus Workbench**

Metadata

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.) ...	<u> (utterance)
paralinguistic event	description	<incident>
sentence	-	<s>

Example

debateRole="speaker"

group="SRC"

name="Laurence Dumont"

nominationParl="Vice-Président"

speechType="MOTION_RP_2_30"

Linguistic annotation (1)

Bonsai + Talisman NLP pipeline for French

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

Linguistic annotation (2)

Lexical Property	Description
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

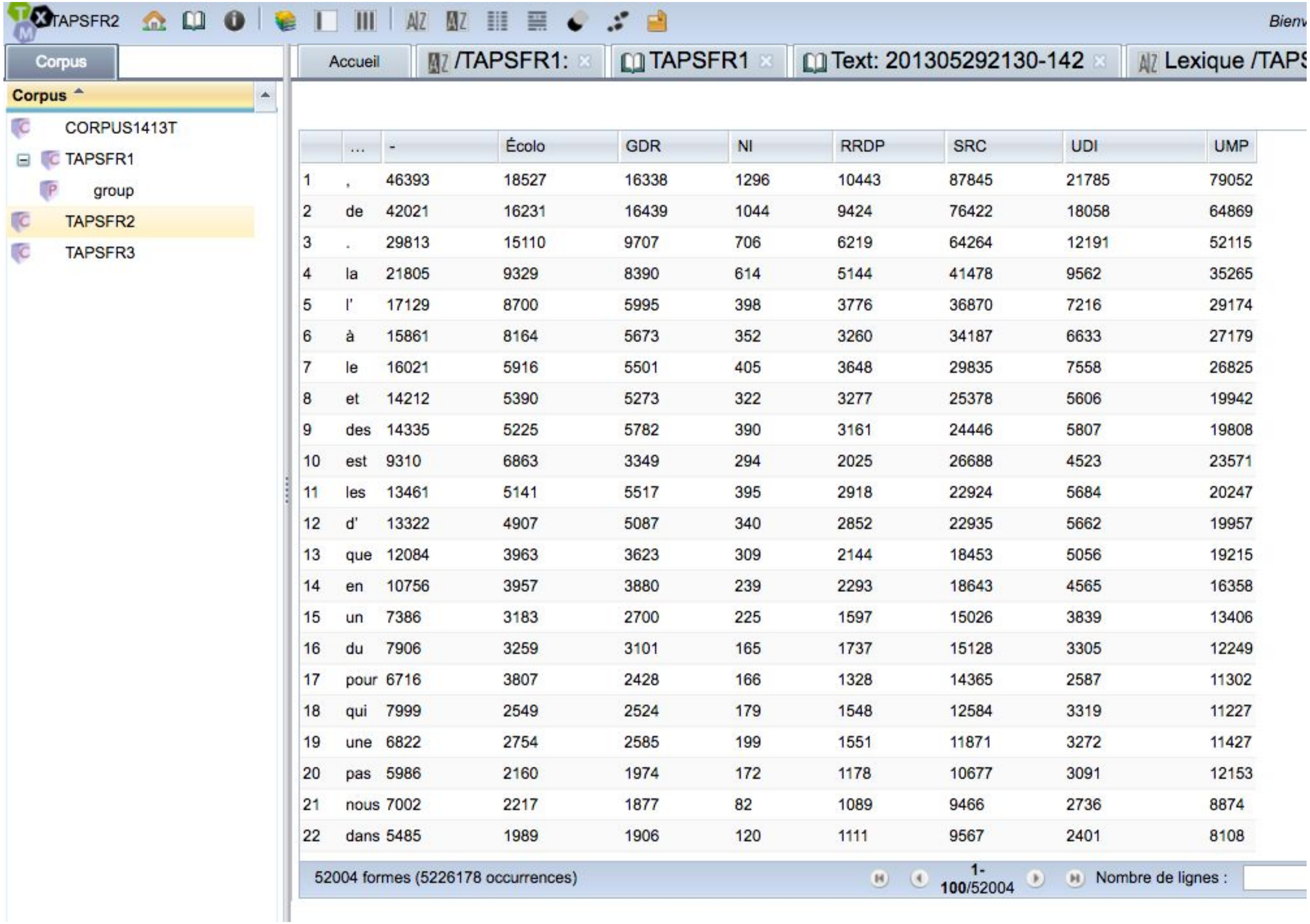
Publication

TAPS-fr is meant to be a ***monitor corpus*** , it will continually be expanded.

Preliminary version is now accessible at http://textometrie.univ-montp3.fr/

Soon: stable, downloadable version on the *Ortolang CLARIN repository* for long term preservation, under the CC BY-NC 4.0 license.

Ortolang >> VLO; parliamentary corpora resource family



	...	-	Écolo	GDR	NI	RRDP	SRC	UDI	UMP
1	,	46393	18527	16338	1296	10443	87845	21785	79052
2	de	42021	16231	16439	1044	9424	76422	18058	64869
3	.	29813	15110	9707	706	6219	64264	12191	52115
4	la	21805	9329	8390	614	5144	41478	9562	35265
5	l'	17129	8700	5995	398	3776	36870	7216	29174
6	à	15861	8164	5673	352	3260	34187	6633	27179
7	le	16021	5916	5501	405	3648	29835	7558	26825
8	et	14212	5390	5273	322	3277	25378	5606	19942
9	des	14335	5225	5782	390	3161	24446	5807	19808
10	est	9310	6863	3349	294	2025	26688	4523	23571
11	les	13461	5141	5517	395	2918	22924	5684	20247
12	d'	13322	4907	5087	340	2852	22935	5662	19957
13	que	12084	3963	3623	309	2144	18453	5056	19215
14	en	10756	3957	3880	239	2293	18643	4565	16358
15	un	7386	3183	2700	225	1597	15026	3839	13406
16	du	7906	3259	3101	165	1737	15128	3305	12249
17	pour	6716	3807	2428	166	1328	14365	2587	11302
18	qui	7999	2549	2524	179	1548	12584	3319	11227
19	une	6822	2754	2585	199	1551	11871	3272	11427
20	pas	5986	2160	1974	172	1178	10677	3091	12153
21	nous	7002	2217	1877	82	1089	9466	2736	8874
22	dans	5485	1989	1906	120	1111	9567	2401	8108

52004 formes (5226178 occurrences)

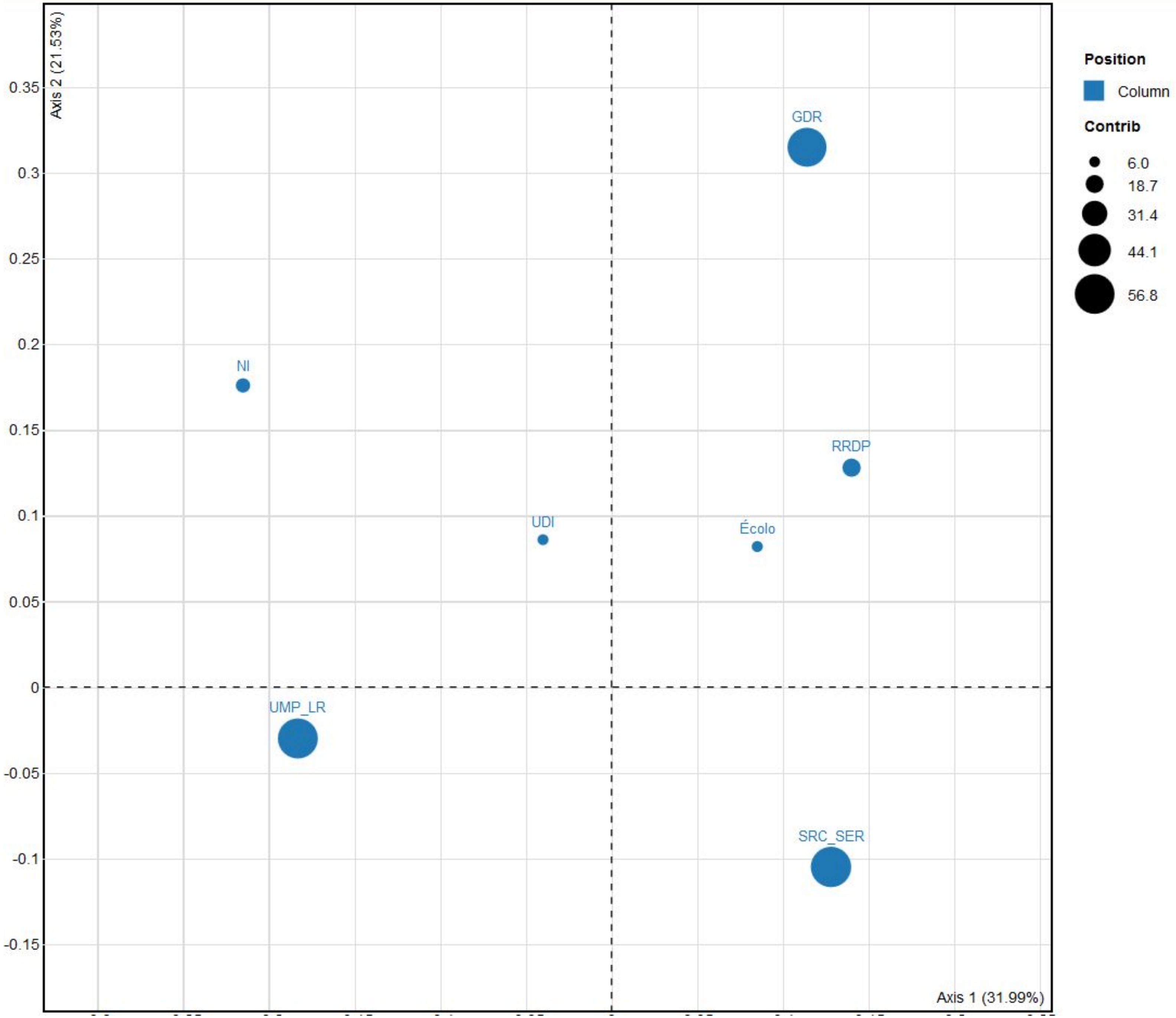
1-
100/52004

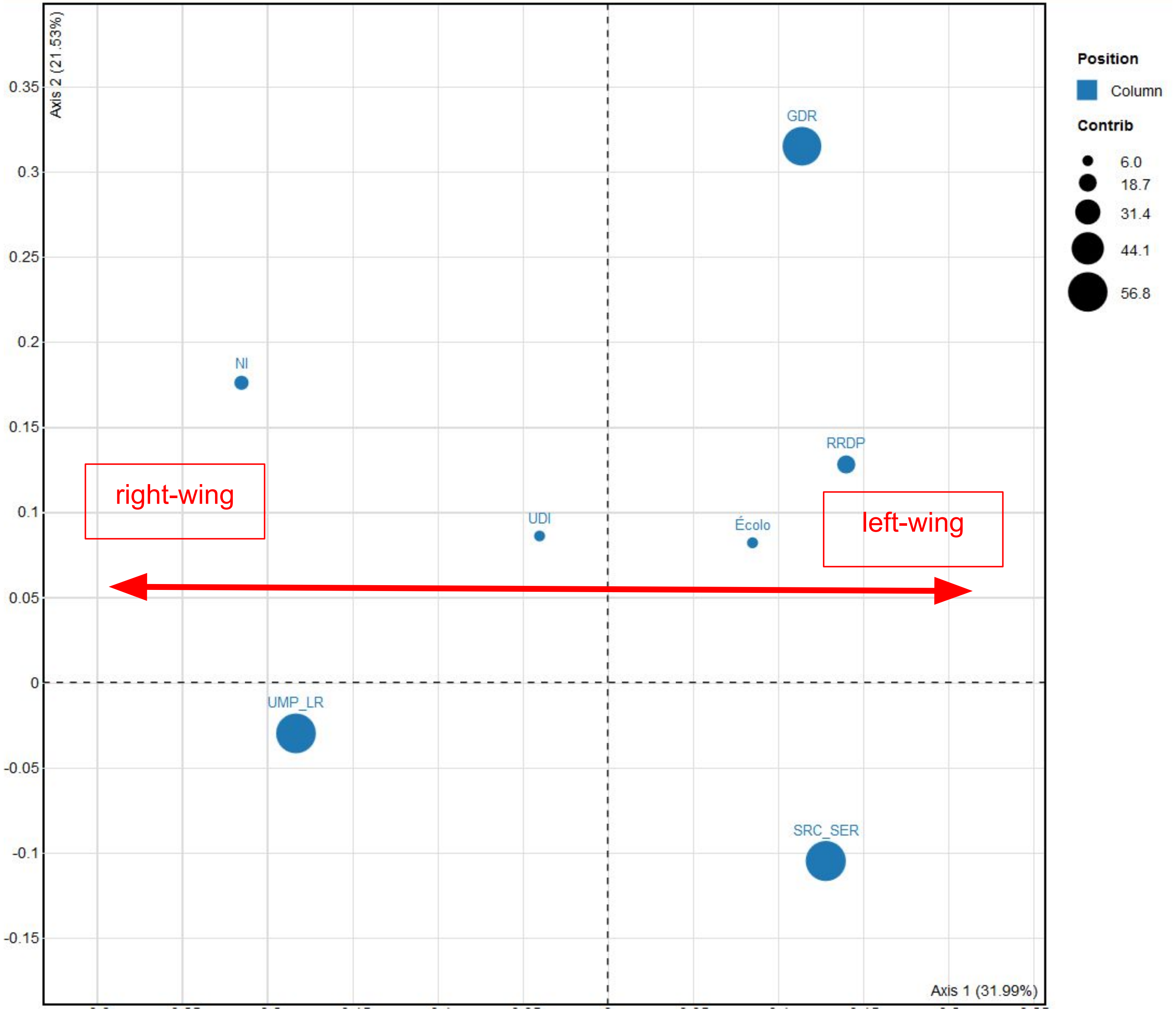
Nombre de lignes :

Exploration - “textometric” approaches

Correspondence Analysis (CA) is a useful technique providing a condensed view of divergences relating to samples (resulting from a partition in the corpus) and countable linguistic features (e.g. lexical items).

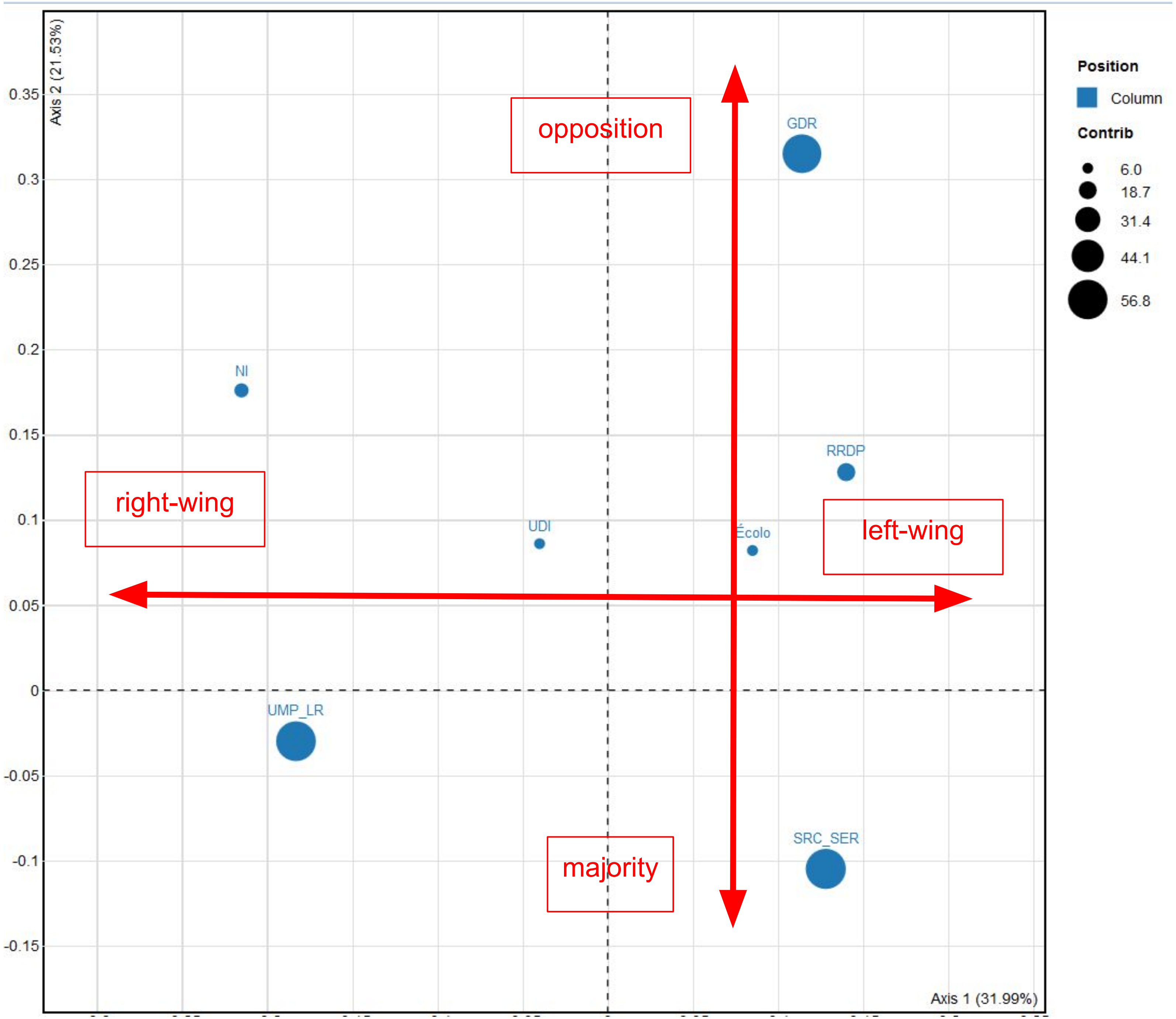
Here is an example of a CA plot based on a **partition by political group**


Axis 2 (21.53%)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
GDR
NI
RRDP
UDI
Écolo
UMP_LR
SRC_SER
Axis 1 (31.99%)
Position
Column
Contrib
6.0
18.7
31.4
44.1
56.8



right-wing

left-wing



opposition

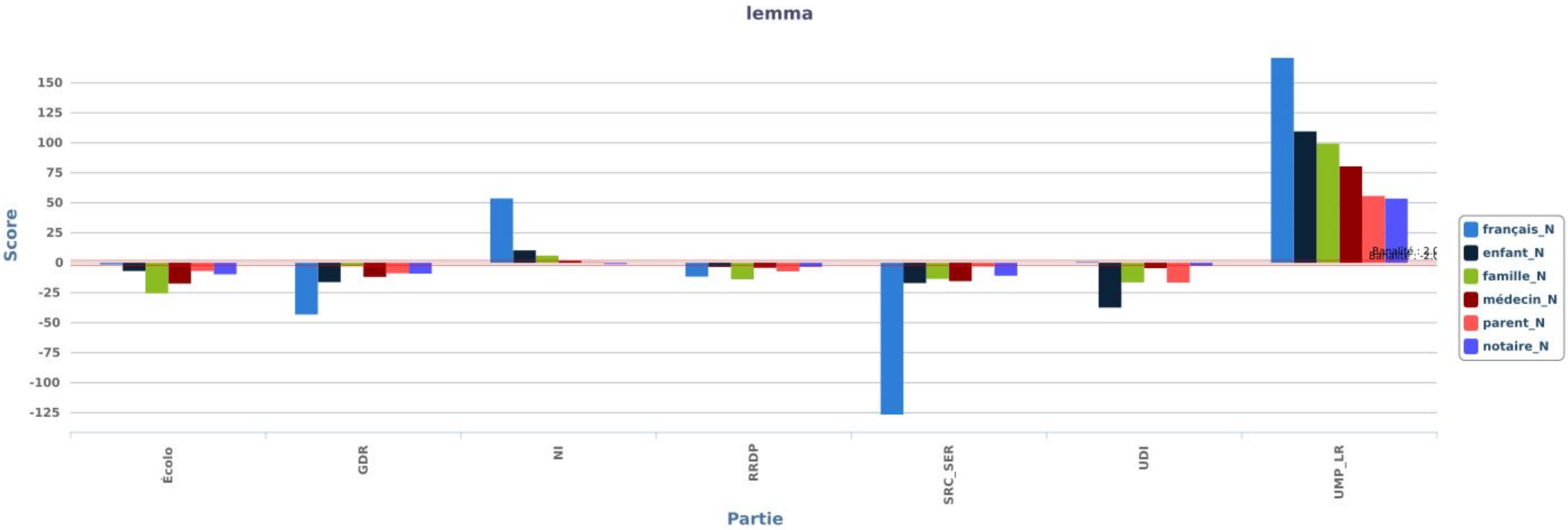
right-wing

left-wing

majority

Exploration - specificities

It is possible to extract the most characteristic nouns specific to the discourse of a given parliamentary group.

lemma
150
125
100
75
50
25
0
-25
-50
-75
-100
-125
Score
Banalité : 2.0
Banalité : -2.0
Écolo
GDR
NI
RRDP
SRC_SER
UDI
UMP_LR
Partie
français_N
enfant_N
famille_N
médecin_N
parent_N
notaire_N