JADT 2018
INTERNATIONAL CONFERENCE ON STATISTICAL ANALYSIS OF TEXTUAL DATA

*On the **phraseology of spoken French**: initial salience, prominence and lexicogrammatical recurrence in a prosodic-syntactic treebank Rhapsodie*

Maria Zimina, Nicolas Ballier
EA 3967 CLILLAC-ARP, Université Paris Diderot – USPC

Outline

- Objectives: explore prosodic dimensions of phraseology
- State-of-the art research
- The *Rhapsodie* speech dataset
- Textometric analysis with ***Le Trameur*/*iTrameur***
- Contrastive analysis across speech genres:
  - Lexicogrammatical recurrence and initial prosodic salience
  - Initial prosodic salience and final prominence
- Conclusions
- Future work

Objectives

- Explore **prosodic dimensions of phraseology**
  - Reveal the link between the "marked status" as a +phrase/expression/formulaic expression etc. and prosodic constituents
- Analyse phraseology using **richly annotated corpora** (morpho-syntactic, syntactic, macro-syntactic and prosodic annotations)

State-of-the-art research

- Nespor and Vogel, 2007; Lin, 2013:
  - Prosodic analysis of phraseological units attested in speech-to-text transcription
  - Prosodic constituents are not explored

- Non-congruency of the recurrence of prosodic features (such as prominence), and traditional phraseological units recovered from transcribed speech data

Rhapsodie speech data: prosodic annotation (Lacheret *et al.*, 2014)

IPE	que vous soyez devenue une vedette vous étiez normalement entrainée																–
IPA	que vous soyez devenue une vedette vous étiez normalement entrainée																
RG	que vous soyez devenue					une vedette			vous étiez			normalement			entrainée		
MF	kvuswajedəvny					ynvədɛt			vuzetje			nɔr	malmã		ãtrene		
syllable	kvu	swa	je	dəv	ny	yn	və	dɛt	vu	ze	tje	nɔr	mal	mã	ã	tre	ne
Prom	0	0	0	0	W	0	0	W	0	0	W	S	0	0	0	0	S

- **57 short samples of spoken French** (~ **5 minutes long**), orthographically and phonetically transcribed (~ **33,000 words**)
- Designed to investigate the **prosody/syntax/discourse interface** across several **discourse types** and **speaking styles**
- Freely available from ***www.projet-rhapsodie.fr***
- More than **60 annotation layers** (morpho-syntactic, syntactic, macro-syntactic and prosodic features)

The *Rhapsodie* prosodic structures

Intonational Periods (IPE)

Intonational PAckages (IPA)

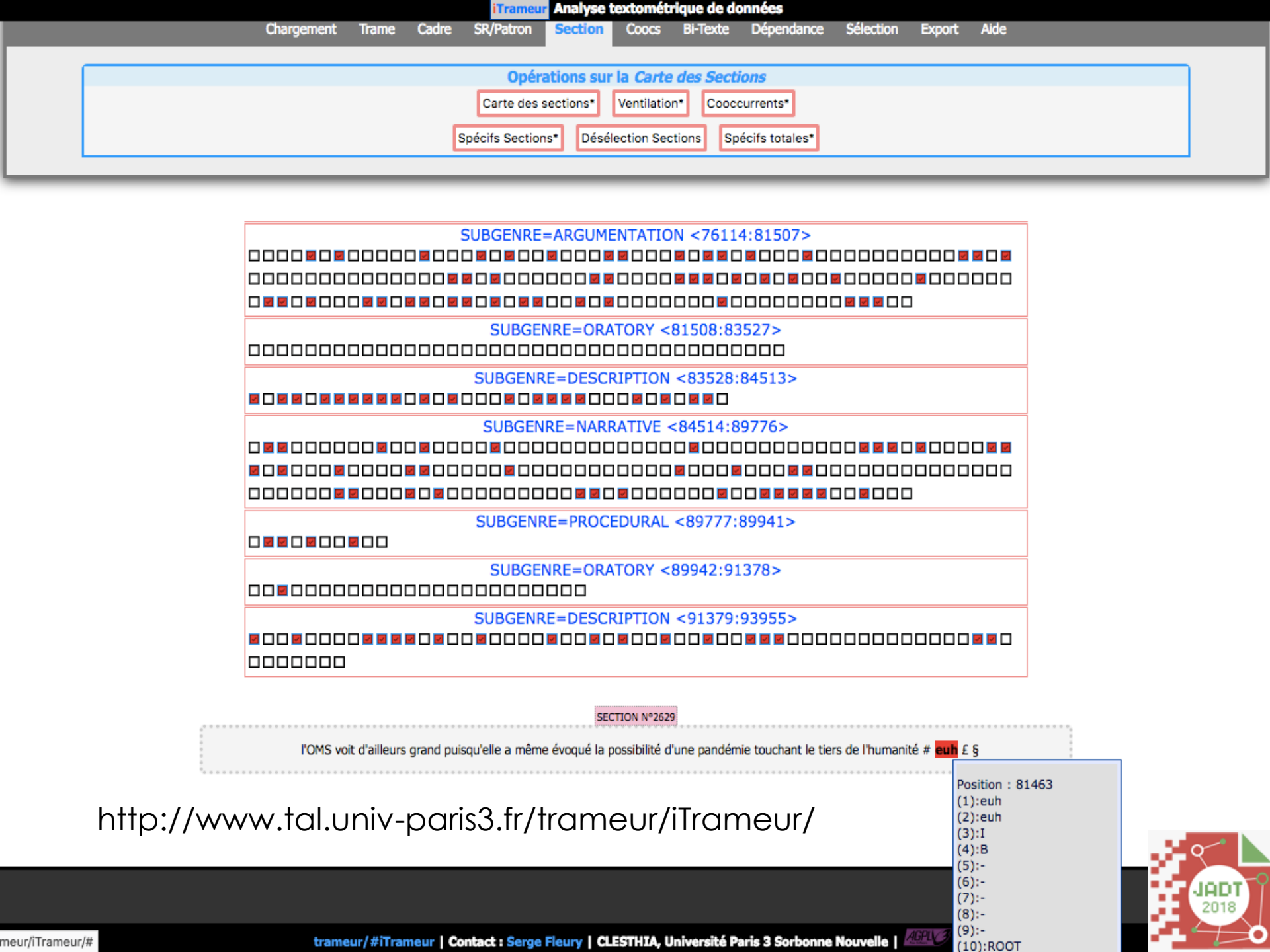
Rhythmic Groups (RG)

Metrical Feet (MF)

Syllables (with Prominence levels: **O**: non-prominent , **S**: strong, **W**: weak)

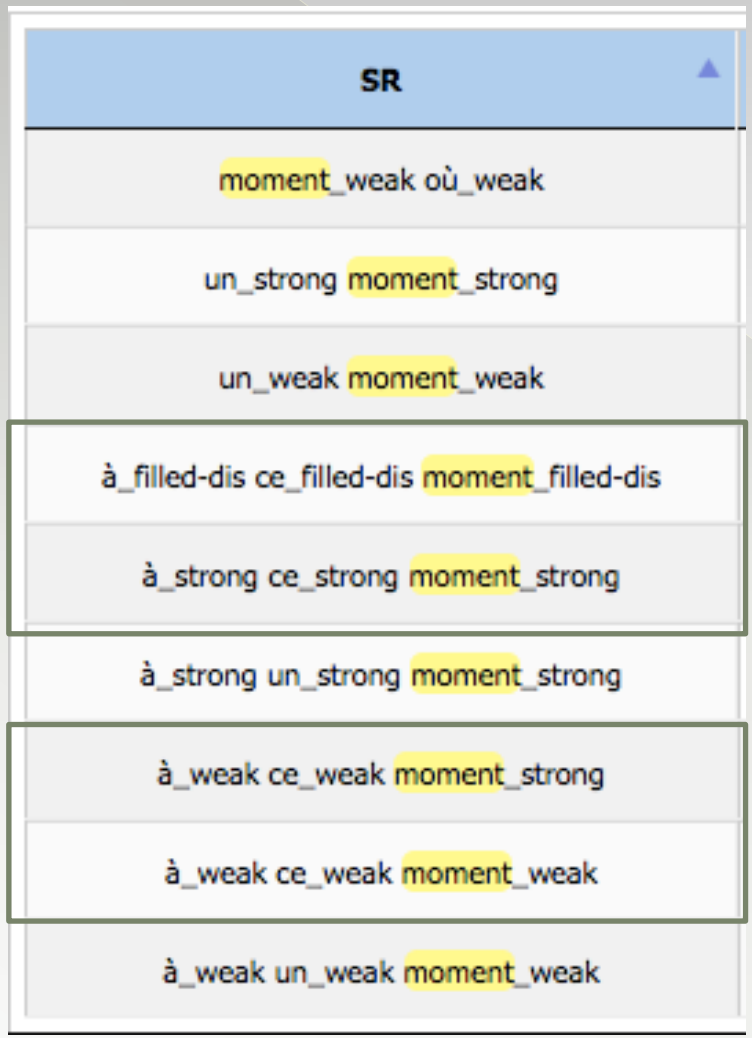
Textometric base file (S. Fleury)

```
1   forme  euh I  euh B  -  -  -  -  -  ROOT  ROOT  -  -  -  -  O  B  O  O  B  O  O  O  O  O  O  O  U  O  0  0  -  92.90196761259227
2   delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
3   forme  bon I  bon B  -  -  -  -  -  ROOT  ROOT  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  U  O  0  0  -  93.61142792134571
4   delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
5   forme  pour  Pre pour  B  -  -  -  -  -  AD(33)  AD(33)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  0  0  -  93.5580384
6   delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
7   forme  aller  V  aller  B  infinitive  -  -  -  -  DEP(5)  DEP(5)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  S  W  -  93
8   delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
9   forme  du  Pre+D  de+le  B  -  -  -  sg  masc  OBJ(7)  OBJ(7)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  0  0  -  93.914
10  delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
11  forme  CRDT  N  CRDT  B  -  -  -  sg  masc  DEP(9)  DEP(9)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  W  0  -  89.883
12  delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
13  forme  à  Pre à  B  -  -  -  -  -  OBL(7)  OBL(7)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  0  0  H  90.26366309887754
14  delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
15  forme  la  D  le  B  -  -  -  sg  fem  DEP(17)  DEP(17)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  0  0  H  92.07666113239492
16  delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
17  forme  gare  N  gare  B  -  -  -  sg  fem  DEP(13)  DEP(13)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  0  S  -  88.1712225
18  delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
19  forme  euh I  euh B  -  -  -  -  -  ROOT  ROOT  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  U  O  0  _  H  86.44444171734648
20  delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
21  forme  de  Pre de  B  -  -  -  -  -  DEP(17)  DEP(17)  -  -  -  -  O  I  O  O  I  O  O  O  O  O  O  O  O  O  0  0  -  88.80414647402179
22  delim  BLANK  BLANK  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
23  forme  Grenoble  N  Grenoble  B  -  -  -  sg  masc-fem  DEP(21)  DEP(21)  -  -  -  -  O  I  O  O  L  O  O  O  O  O  O  O  0  0  0  0
```



http://www.tal.univ-paris3.fr/trameur/iTrameur/

Prominence and traditional phraseology (*Rhapsodie*)



SR
moment_weak où_weak
un_strong moment_strong
un_weak moment_weak
à_filled-dis ce_filled-dis moment_filled-dis
à_strong ce_strong moment_strong
à_strong un_strong moment_strong
à_weak ce_weak moment_strong
à_weak ce_weak moment_weak
à_weak un_weak moment_weak

j_weak'_weak étais_weak communiste_strong **à_strong ce_strong moment_strong-_strong là**_strong # £

mais_weak **à_weak ce_weak moment_weak-_strong là_strong** il_strong y_strong a_strong une_strong dame_weak un_weak peu_weak plus_weak âgée_ …

Prosodic phraseology: quantitative analysis (first results)

- Intonational PEriods (IPE), segments of speech with distinctive pitch and rhythm contours, are strongly related to spoken formulaic language (Zimina and Ballier, 2017)
  - new insights into the observation of the functions of formulaic expressions in speech

‘Lexicogrammar’ approach
(Gledhill et al., 2017)

- Explore the ways in which **prosodic features** may correlate with extended **lexicogrammatical patterns** with a particular **register** or **discourse function** (Gledhill *et al.*, 2017)

  - § et donc euh < **c'est pour** ça **qu'**aujourd'hui je suis en italien en XXX /…/ >
  - § c'est-à-dire § ouais § un mois <**c'est pour** ça **que** ça s'appelle radio Timsit /…/ >
  - § **c'est pour** cela **que** je tenais à vous rencontrer la veille de notre fête /…/

Textometric analysis

Textometric base file (more than **60 annotation levels**)

Processing multiple annotation levels with ***Le Trameur*/*iTrameur*** (S. Fleury)

Automatic re-annotation: *joining POS, prosodic constituents (BILOU) & final prominence*

Repeated segments on joined annotations

Characteristic elements (discourse types, speaking contexts, etc.)

RESULTS

Initial prosodic salience and lexicogrammatical recurrence

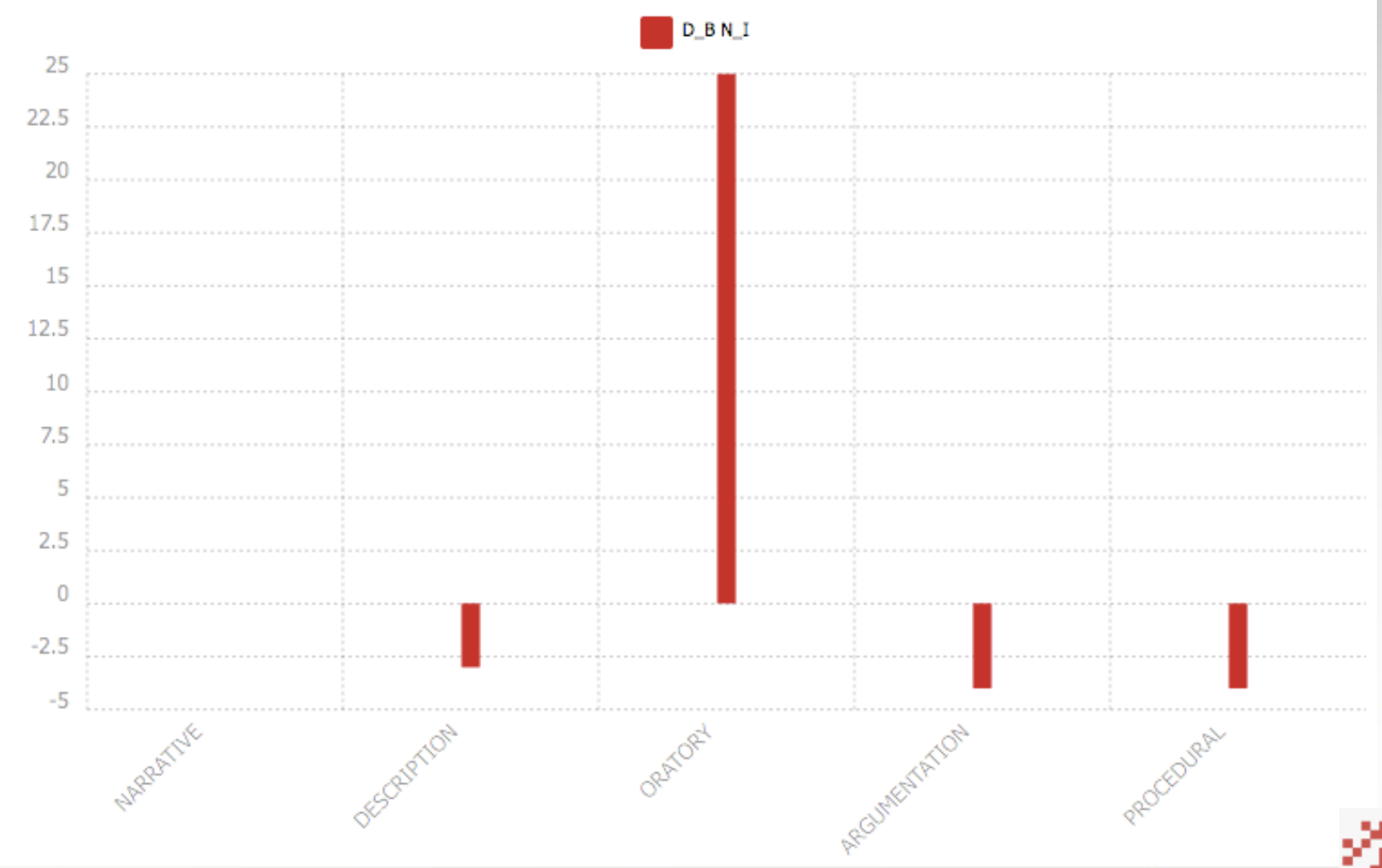
Systemic combination of initial salience and final prominence

Lexicogrammatical recurrence and prosodic salience (B: beginning of Intonational Periods: IPE)

SR	Fq	Lg
Cl_B V_I	260	2
D_B N_I	134	2
J_B Cl_I	114	2
Pre_B D_I	112	2
Cl_B Cl_I	102	2
J_B Adv_I	72	2
Adv_B Cl_I	71	2
J_B I_I	70	2
Cl_B Cl_I V_I	69	3
I_B I_I	65	2

Most **frequent initial pivots** at the beginning of IPE

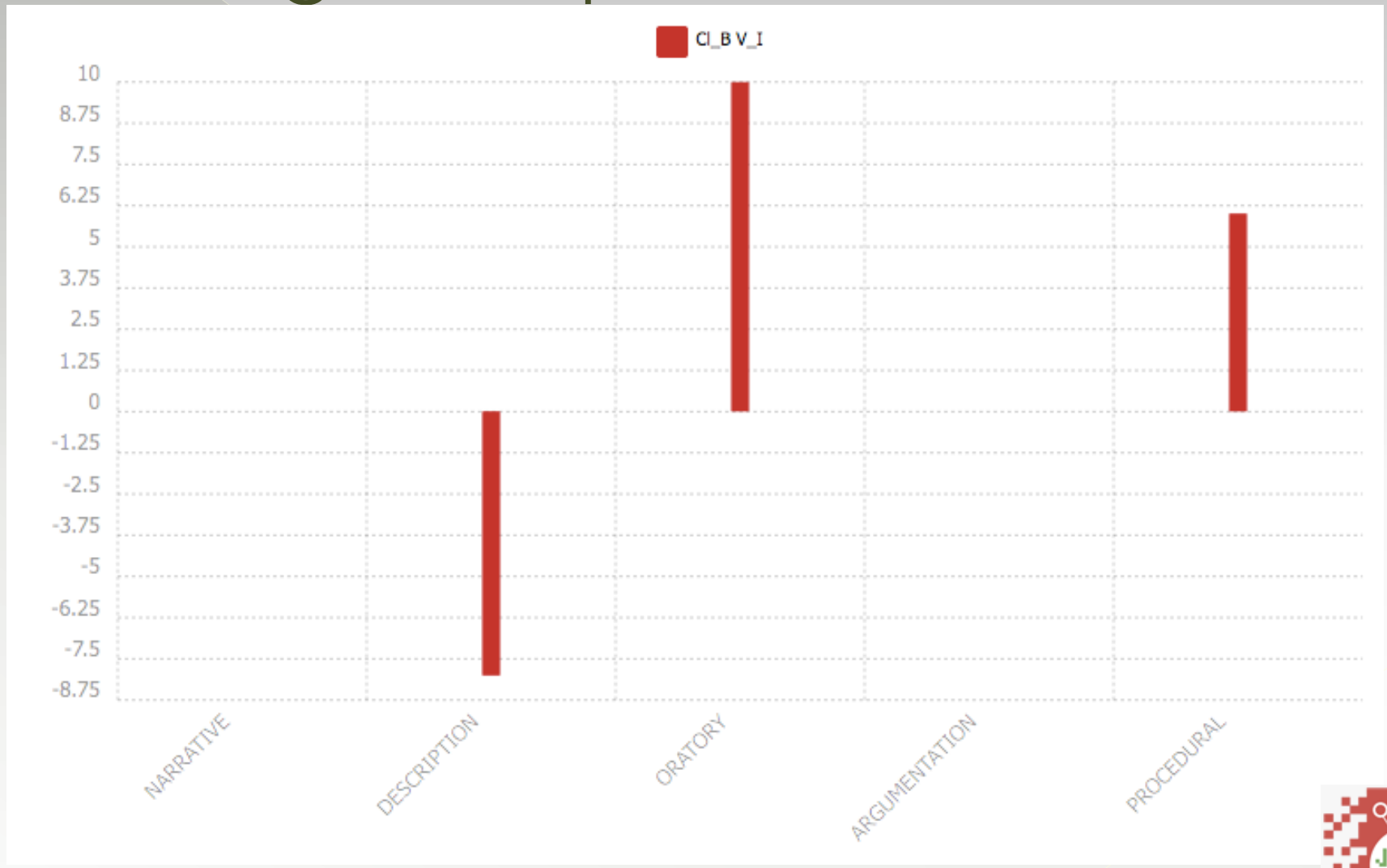
Det +Noun in the position of strong initial prosodic salience



Det + Noun (theme selection)

N°	Partie	Contexte Gauche	Pôle	Contexte Droit
22	ORATORY	en ces heures difficiles nous ressentons # profondément la fragilité des choses #	la	précarité de ce qui nous semblait acquis # £ § nous voyons combien # tout
23	ORATORY	'importance du rôle de l'État dans notre société #	un	État sur lequel pèsent des responsabilités essentielles # le service public # la
24	ORATORY	dans notre société # un État sur lequel pèsent des responsabilités essentielles #	le	service public # la sécurité # la solidarité # un État # auquel il appartient
25	ORATORY	même communauté # et d'être # responsables les uns des autres # £ §	la	France blessée # veut se retrouver # rassemblée et fraternelle # £ § parce que nos
26	ORATORY	parler leur coeur # je voudrais dire # merci à tous les Français # £ §	ce	soir nous vivons ensemble # un moment fort # et singulier # £ § ce qui
27	ORATORY	dire # merci à tous les Français # £ § ce soir nous vivons ensemble #	un	moment fort # et singulier # £ § ce qui paraissait très lointain et qui
28	ORATORY	éducation les conditions de vie # pour les libertés la vie démocratique #	la	situation des femmes les solidarités # mais aussi # siècle d'horreur
29	ORATORY	tragédies de convulsions # qui a vu deux guerres mondiales # le goulag #	les	dictatures totalitaires # et la Shoah # £ § mais ce soir ce qui importe
30	ORATORY	'est l'avenir # notre avenir # celui de nos enfants # £ §	le	progrès va se poursuivre avec ses hésitations # avec ses limites # que
31	ORATORY	au clonage # £ § de même dans le domaine de l'environnement #	les	peuples ne veulent plus # que la course à la produc # épuise

Clitic +Verb in the position of strong initial prosodic salience

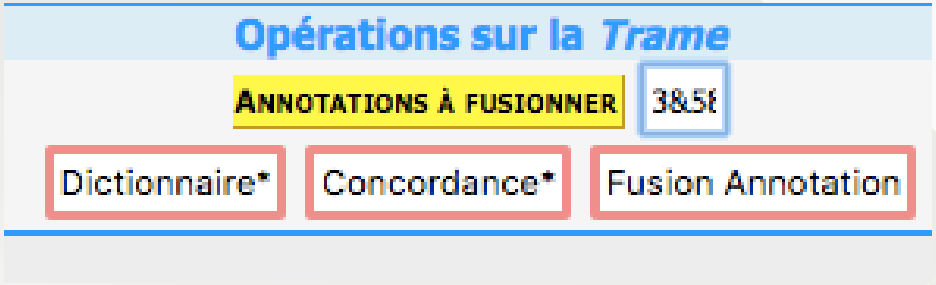


Clitic +Verb (performatives OR instructions)

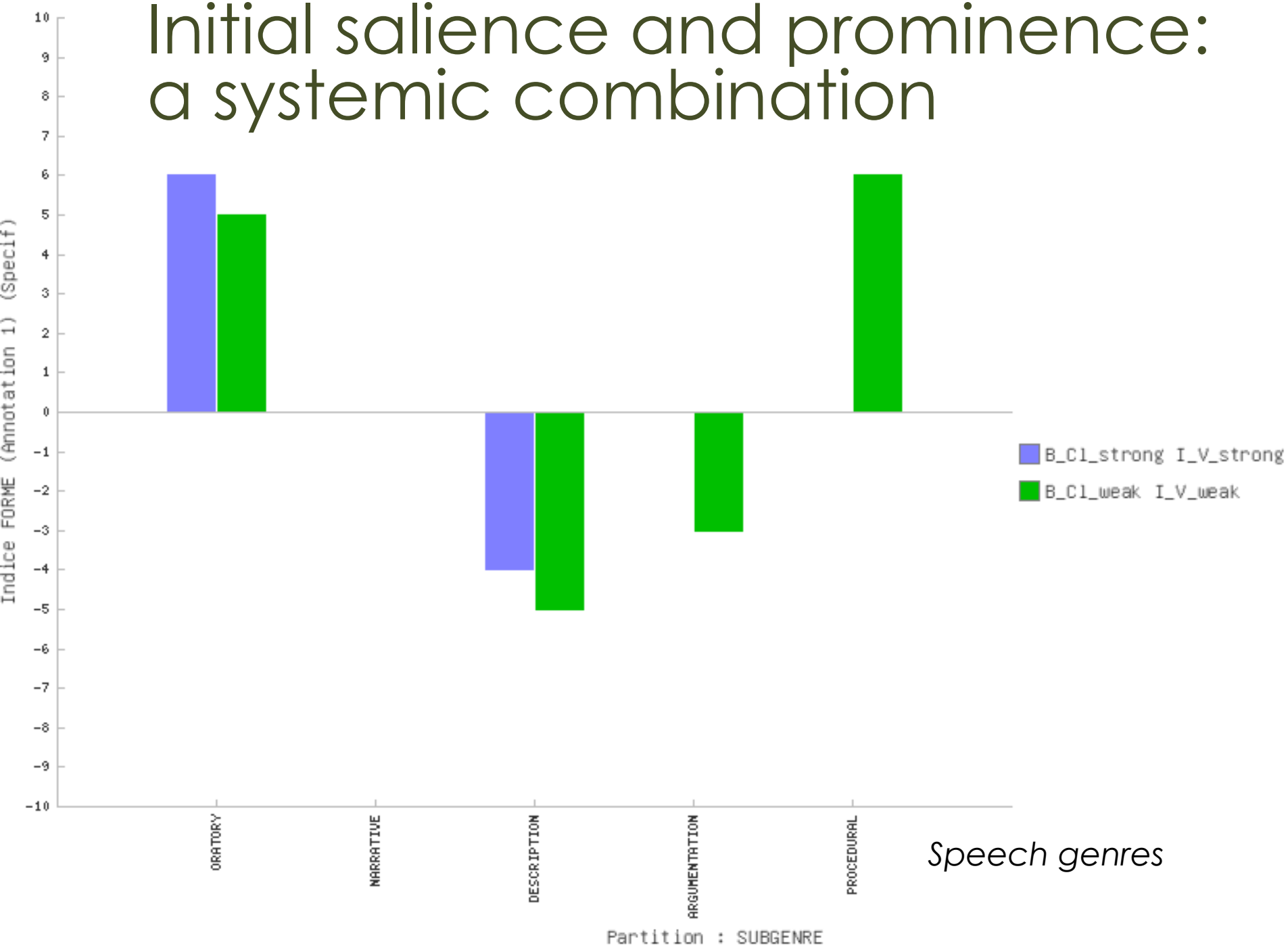
N°	Partie	Contexte Gauche	Pôle	Contexte Droit
247	ORATORY	fête nationale # soit aussi celle de l'Europe # £ § demain #	ce	sont les vingt-six drapeaux de nos partenaires européens # qui
248	ORATORY	tenir son rang # £ § les bases d'une défense européenne existent # £ §	il	faut les faire grandir # en quittant le terrain des mots # pour
249	ORATORY	de l'action # £ § demain davantage qu'aujourd'hui #	je	souhaite que l'Europe # soit capable d'assurer sa
250	ORATORY	plus en plus dangereux # £ § l'Afghanistan # le Proche-Orient #	je	connais la somme de courage et d'abnégation # que requiert
251	ORATORY	requiert l'accomplissement de vos missions dans un tel contexte # £ §	je	sais également ce que cela signifie pour vos familles # que je
252	ORATORY	souvent confrontées à l'absence # et parfois l'angoisse # £ §	je	sais aussi hélas le lourd tribut # payé par certains de vos
1	PROCEDURAL	donc tou~ tou~ toujours Saint-Jean-de-Maurienne # £ §	tu	passes un autre rond-point £ § donc toujours tout droit # £ § après
2	PROCEDURAL	toujours tout droit # £ § après tu montes une grande grande ligne droite # £ §	tu	passes devant la piscine # £ § euh il y a un stade aussi
3	PROCEDURAL	-point £ donc là tu rentres dans le centre-ville # £ §	tu	arrives à un rond-point £ § c'est donc à
4	PROCEDURAL	tu descends £ § en fait c'est une route qui descend # £ §	tu	passes devant les pompiers # £ § et euh ensuite premier euh &

Prosodic salience and prominence of final syllables

- **Combined annotation** levels:
  - (1) IPE structure with *BILOU* tags (*Beginning*, *Inside*, *Last, Unit*-length and *Outside*)
  - (2) POS tags
  - (3) final prominence: *strong*, *weak*, *pause_* and % (inaudible or non-transcribed due to overlap)

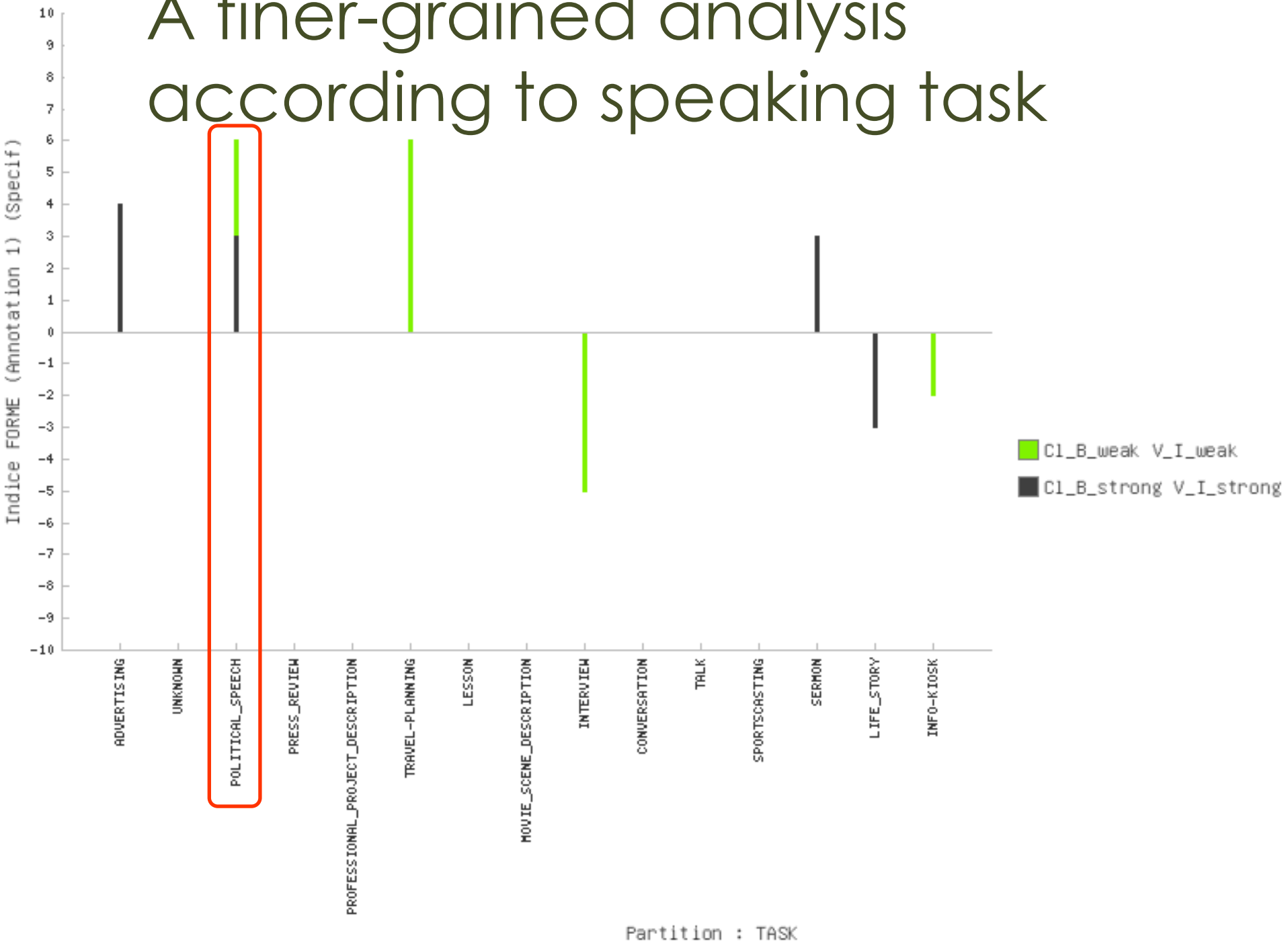


Initial salience and prominence: a systemic combination



Speech genres

A finer-grained analysis according to speaking task



First results

- At the **beginning (B) of IPEs**, the **final prominence** of the pivot **CI + V** is either *weak weak* or *strong strong*
- In political speech, pragmatic strategies influence the choice of a *weak weak* (+06) or of a *strong strong* (+03) sequence to realize specific discourse functions, for example:

*…reculer la pauvreté # § **ce|strong sera|strong** tout le sens du combat de la France… (focus)*

*…ces valeurs # § en les faisant vivre # **nous|weak serons|weak** plus forts pour aborder les temps qui viennent… (fonction performative)*

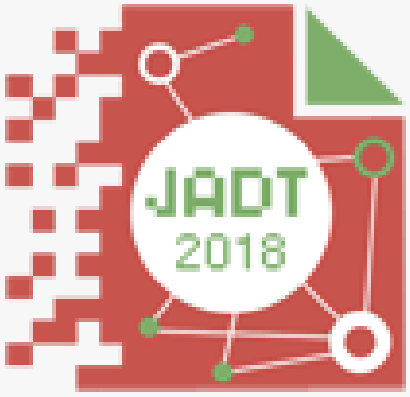
- The strong prominence of **CI + V** also corresponds to emphatic realisations at the beginning of IPE in sermons (+03) and advertising (+04)

Conclusions

- Limited distribution of POS categories in the initial position: Intonational PEriod (IPE)
- Recurrent patterns reflected by such sequences as *"je salue"*, *"elle souhaite"*, *"il faut"*, *"on continue"* are not unlike the stable lexicogrammatical patterns that can be observed in written data
- Relevance of prosodic prominence (stressed syllables) for the distinction of speech genres
- Initial characteristic distributions with specific prosodic characteristics correspond to communicative needs (interactions, uptakes, speaking turns, etc.)

Future work

- After preliminary explorations:
  - other layers of annotations
  - other layers of granularity: Intonational PAckage (IPA), Rhythmic Group (RG), etc.
  - other variables (channel, planning type, event structure: monologal *vs.* dialogal tasks)
- Humming (erasing lexical contents, only keeping the melody): identifiable characteristic signals of collocations



JADT 2018

INTERNATIONAL CONFERENCE ON STATISTICAL ANALYSIS OF TEXTUAL DATA

Thank you

- mzimina@eila.univ-paris-diderot.fr
- nicolas.ballier@univ-paris-diderot.fr

References

- Dorna, A. (1995). Les effets langagiers du discours politique. *Hermès, La Revue* 1995/2 16, 131–146.
- Fleury, S. (2013). *Le Trameur. Propositions de description et d'implémentation des objets textométriques*. Sorbonne nouvelle – Paris 3, http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf
- Gledhill C., Patin S., Zimina M. (2017). Identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. *CORPUS* 17, pp.
- Granger, S. (2005). Pushing back the limits of phraseology. How far can we go? In: Cosme, C., Gouverneur, C., Meunier, F., Paquot, M. (eds.): *Proceedings of PHRASEOLOGY 2005. An Interdisciplinary Conference*, Université Catholique de Louvain, Louvain-la-Neuve, pp. 165–168.
- Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J-P., Obin, N., Pietrandrea, P., Tchobanov, A. (2014). *Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.
- Lin, Ph. M.S. (2013). The prosody of formulaic expression in the IBM/Lancaster Spoken English Corpus International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics* 18(4), 561–588.
- Nespor, M., Vogel, I. (2007). *Prosodic Phonology*. Berlin. Mouton De Gruyter.
- RHAPSODIE Homepage, http://www.projet-rhapsodie.fr
- Sitri, F., Tutin, A. (dir.) (2016). Phraséologie et genres de discours. Patrons, motifs, routines. *LIDIL* 53.
- Yoo, H-Y, Delais-Roussarie, E. (eds.) (2009). *Actes de la conférence Interface Discours & Prosodie (IDP 2009)*, Paris, France, http://makino.linguist.jussieu.fr/idp09/actes_fr.html
- Zimina, M., Ballier, N. (2017). Intonational PEriods (IPE) and Formulaic Language: A Genre-based Analysis of a French Speech Database. *Proceedings of Europhras 2017 Conference: Computational and Corpus-based Phraseology: Recent Advances and Interdisciplinary Approaches*, London, http://www.tradulex.com/varia/Europhras2017-II.pdf