Intonational PEriods (IPE) and Formulaic Language: A Genre-based Analysis of a French Speech Database

Maria ZIMINA-POIROT, **Nicolas BALLIER**
Univ Paris Diderot, Sorbonne Paris Cité,
CLILLAC-ARP, EA3967, 75013, Paris, France

maria.zimina@eila.univ-paris-diderot.fr
nicolas.ballier@univ-paris-diderot.fr

université PARIS DIDEROT PARIS 7
USPC Université Sorbonne Paris Cité

Plan

- **Phraseology** and **prosody**
- The ***Rhapsodie*** speech dataset
  - Perception-driven prosodic annotation (methodology)
  - Prosodic structures (hierarchy)
- **Formulaic expressions and prosodic constituents (exploratory work)**
  - 60 annotation layers (prosody, micro- and macro-syntax)
  - Prosodic features and extended 'lexicogrammatical' patterns
  - Textometric analysis
- First conclusions
- Future work

Phraseology and prosody

"If most of the formulaic expressions we know have been acquired from and are used in speech, the phonological representation of formulaic expressions should, in theory, play a fundamental role in the lexical storage and retrieval." (Lin, 2013)

- **Lin, Ph. M.S.:** The prosody of formulaic expressions in the IBM/Lancaster Spoken English Corpus. *International Journal of Corpus Linguistics* 18 (4), 561–588 (2013).
- **Aston, G.:** Learning phraseology from speech corpora. In: Leńko-Szymańska, A., Boulton, A. (eds.) *Multiple Affordances of Language Corpora for Data-driven Learning (Studies in Corpus Linguistics 69)*, pp. 63–84. John Benjamins, Amsterdam-Philadelphia (2015).

Examples of corpus-based approaches

- Expert-based annotations of stresses and tones: **Lin**
- Manual annotation of *TED Talks*: **Aston**

- BUT queried on the basis of textual sequences:
- Prosodic features are not exploited as such

The *Rhapsodie* speech database

- Composed of **57 short samples of spoken French** (approximately **5 minutes long**), orthographically and phonetically transcribed (approximately **33,000 words**).

- Designed to investigate the **prosody/syntax/discourse interface** across several **discourse types** and **speaking styles** (oratory, narrative, description, argumentation, procedural; interactive, public and private; semi-interactive and non-interactive, etc.)

- Freely available from ***www.projet-rhapsodie.fr***
- More than **60 annotation layers** (morpho-syntactic, syntactic, macro-syntactic and prosodic features)

The *Rhapsodie* methodology for 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

(1) Manual annotation of relevant **perceptual prosodic events**.

(2) Automatic characterization of the **prosodic constituents** based on this manual annotation.

(3) Automatic stylization of **melodic contours** and annotation of **tones** associated with the prosodic constituents.

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)

Perception-driven prosodic annotation and phraseology: exploratory work

- Exploration of the link between formulaic expressions and prosodic constituents (ready-made speech blocks)

- Contribution of the initial structure of IPE macro-units and **recurrent patterns of initial prosodic salience** to the perception of **formulaic language**

- Quantitative analysis of the recurrent prominences observable after speech breaks (***Le Trameur***: *http://www.tal.univ-paris3.fr/trameur)*

Textometric base file
(Fleury & Zimina, COLING 2014)

Thread/ Frame (extract)

```
<item type="delim" pos="46"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="47"><f>lance</f><c>B_V</c><l>lancer</l><a>indicative</a><a>present</a><a>3</a><a>sg</a><a>-</a><a>ROOT</a><a>-</a><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="48"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="49"><f>un</f><c>B_D</c><l>un</l><a>-</a><a>-</a><a>-</a><a>sg</a><a>masc</a><a>DEP(51)</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="50"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="51"><f>appel</f><c>B_N</c><l>appel</l><a>-</a><a>-</a><a>-</a><a>sg</a><a>masc</a><a>OBJ(47)</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="52"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
```

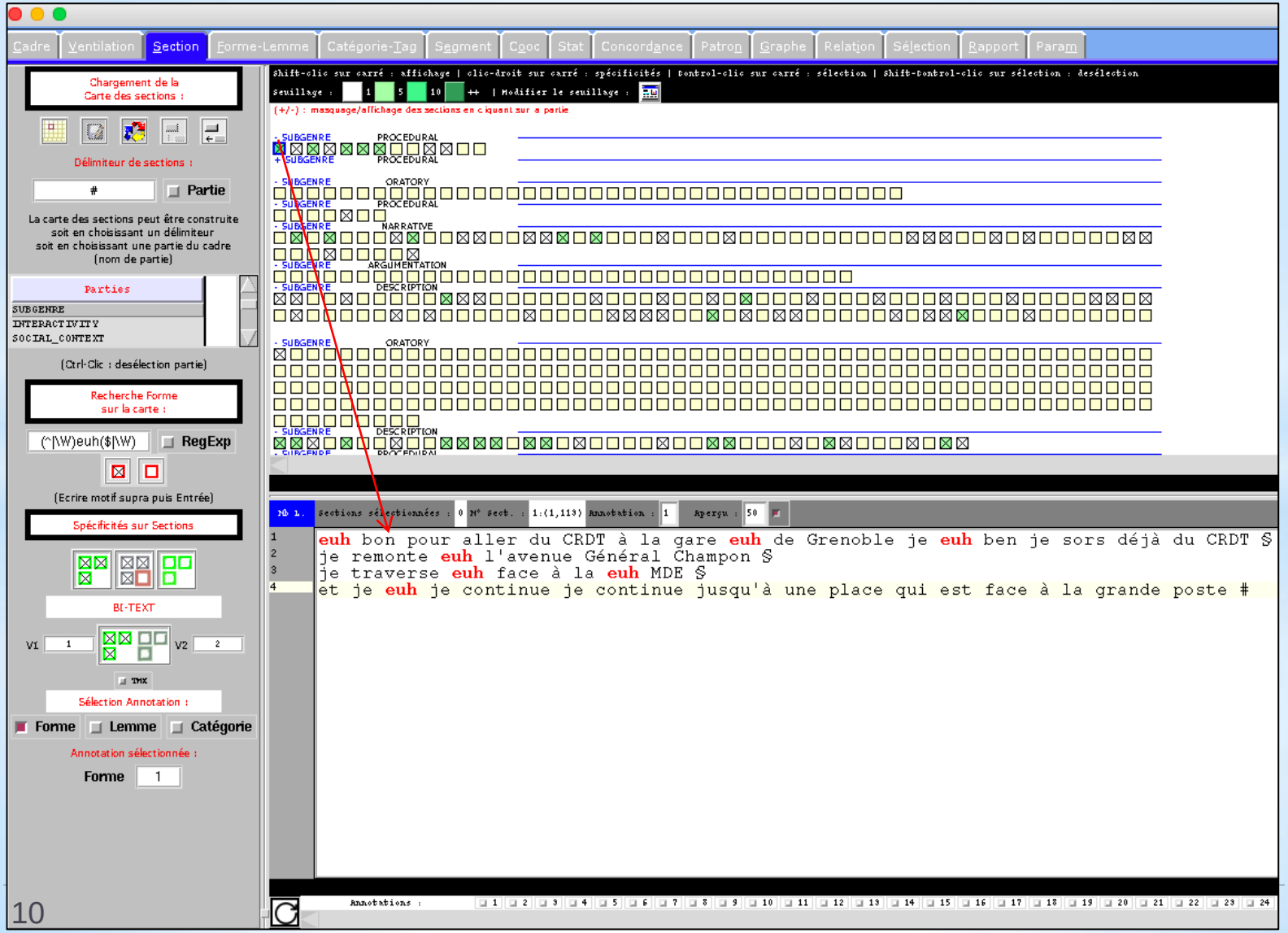
Thread

```
<p n="D2013 " d="58359" f="59730" nd="85" nf="86"/>
<p n="M2006 " d="1" f="2086" nd="1" nf="2"/>
<p n="M0015 " d="40347" f="40514" nd="59" nf="60"/>
<p n="M0022 " d="60079" f="60554" nd="89" nf="90"/>
<p n="M0010 " d="33229" f="33372" nd="41" nf="42"/>
```

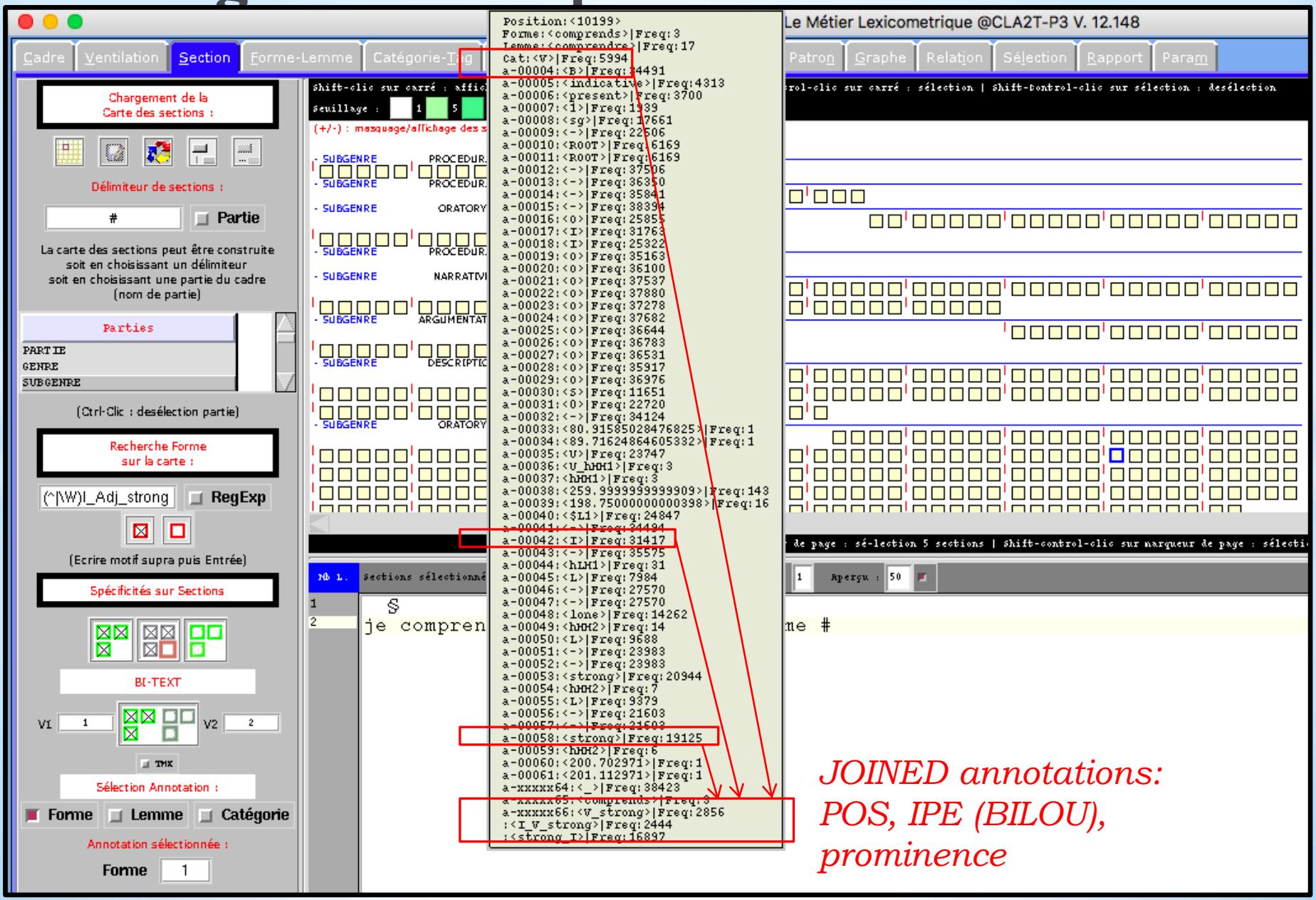
Frame

Final version: **61 annotation levels** (prosody, micro- and macro-syntax).

Le Trameur: the *Rhapsodie* treebank



Working with multiple annotations levels



‘Lexicogrammar’ approach: CLILLAC-ARP

- Predictable and productive sequences of signs called **lexicogrammatical patterns** (corpus studies)
- Composed of permanent ‘**pivotal**’ signs and a more productive ‘paradigm’, these patterns may be discontinuous and may or may not be syntactic constituents (Gledhill *et al.*, 2017)

<*Il y a lieu de* **procéder à** *une évaluation des différentes méthodes de recyclage*>

<*il convient de* **procéder à** *un second examen de toutes les régions du système nerveux qui présentent ces altérations*>

- We aim to explore the ways in which **prosodic features** may correlate with **extended lexicogrammatical patterns**, as well as the extent to which prosody corresponds to patterns which have a particular **register** or **discourse function**.

IPEs: the most frequent POS in the initial position of strongest salience (2, 609 occ.)

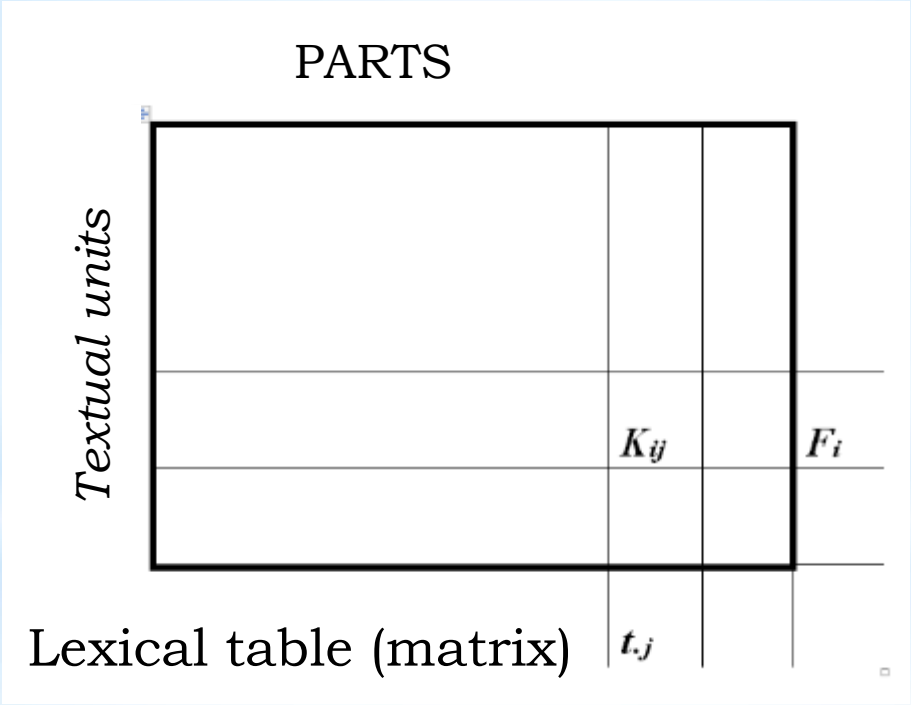
Part-Of-Speech (POS)	Strongest initial prosodic salience	Total of the POS (any position)
Cl (Clitic pronoun)	511 occ.	4 179 occ.
J (Coordinating conjunction)	443 occ.	1 142 occ.
I (Interjection)	439 occ.	1 984 occ.
Adv (Adverb)	287 occ.	2 789 occ.
Pre (Preposition)	238 occ.	3 443 occ.
D (Determiner)	209 occ.	4 080 occ.
V (Verb)	112 occ.	5 994 occ.
Qu (Relative pronoun)	97 occ.	799 occ.
CS (Subordinating conjunction)	74 occ.	729 occ.
N (Noun)	65 occ.	6 317 occ.

IPEs: the most frequent POS recurrences (POS N-grams) in the initial position

POS repeated segment *N*-gram list	Strongest initial prosodic salience (B_IPE)	Total of the N-gram (any position)
CL + V	257 occ.	2 223 occ.
D + N	129 occ.	2 919 occ.
Pre + D	90 occ.	1 112 occ.
J + Cl	77 occ.	164 occ.
Cl + Cl	76 occ.	525 occ.
J + Adv	70 occ.	150 occ.
Cl + Cl + V	69 occ.	479 occ.
J + I	67 occ.	107 occ.
I + I	60 occ.	258 occ.
Pre + D + N	55 occ.	939 occ.

Characteristic elements
(Lebart *et al.*, 1998)

- ***K.*** size of corpus
- F_i frequency of unit ***i*** in entire corpus
- K_{ij} sub-frequency of unit ***i*** in part ***j***
- t_j size of part ***j***

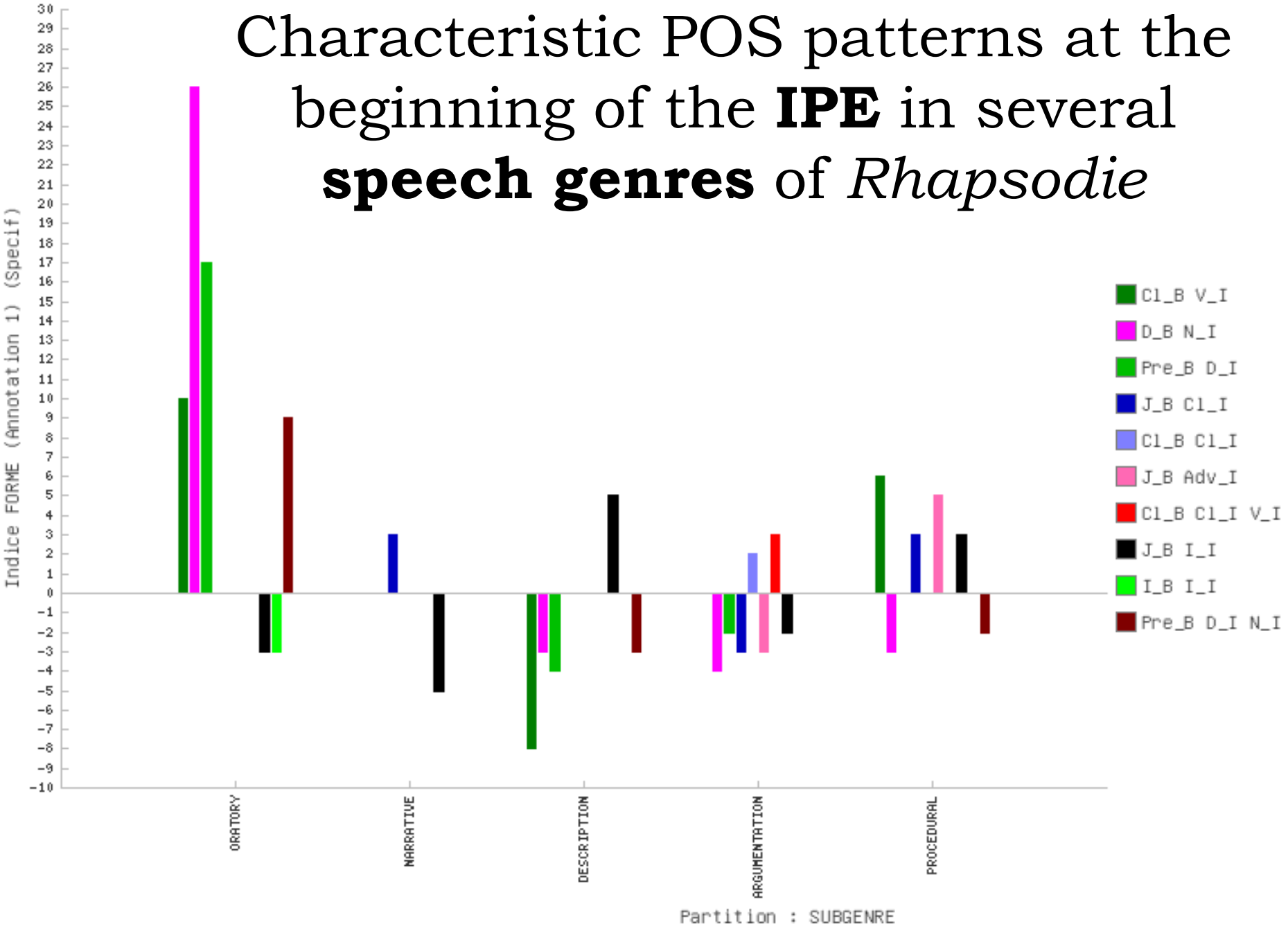


For each cell K_{ij} , we assess the sub-frequency using a *probabilistic model*.

If P_{inf} K_{ij} is smaller that the fixed *threshold*, the unit ***i*** *is a negative characteristic element*.

If P_{sup} K_{ij} is less that the fixed *threshold*, the unit ***i*** *is a positive characteristic element*.

Characteristic POS patterns at the beginning of the **IPE** in several **speech genres** of *Rhapsodie*



Characteristic elements of initial salience in different speech contexts

Oratory: CL + V	**Oratory: D + N**	**Procedural: Cl + V**
• # **je suis** heureux de me retrouver ce soir #	• # **la démocratie** politique et sociale #	• # **on passe** devant le le kiosque à journaux #
• […] est la nation entière qui vous rend hommage # **elle salue** la loyauté #	• # **la France** sera ce que nous voudrons qu'elle soit # une nation unie #	• # **tu vas** tout droit #
• # **il faut** les faire grandir #	• # **le droit** de grève # le droit à l'instruction #	• # **vous continuez** # vous prenez le rond-point tout droit #
• # **je souhaite** que l'Europe #	• # **un moment** fort #	• # **on traverse** la rue #
	• # **l'exigence** de solidarité #	• # **tu descends** toute la pente #
	IPE boundaries	
performatives	***theme-selection***	***instructions***

First conclusions

- Speech boundaries of intonational periods (IPE) can be easily related to the **characteristic repetitions of lexicogrammatical patterns at the beginning of the IPE**.

- The '**pivots**' of these productive patterns have stable lexicogrammatical realizations in the *Rhapsodie* speech dataset (such as the expression of predicates *CL+V* "je salue", "elle souhaite", "il faut", "on continue", etc.).

- **Recurrent initial patterns vary in different social contexts** and reflect speech signals to which speakers and listeners respond in a distinct way (intrinsic experience of language acquisition, specific communicative needs)

Future work

- Integration of **other prosodic characteristics** (tone, pause length, etc.) available in the ***Rhapsodie*** speech dataset (more than 60 annotation layers)
- Inclusion of **other prosodic constituents** (complex and interdependent borders of the prosodic structures)
- Processing capabilities of the annotators in determining IPE boundaries when it comes to formulaic language (**perceptive criteria**)
- Comparison with **other speech data** collections (French, English, etc.)

References

- **Aston, G.:** Learning phraseology from speech corpora. In: Leńko-Szymańska, A., Boulton, A. (eds.) *Multiple Affordances of Language Corpora for Data-driven Learning* (Studies in Corpus Linguistics 69), pp. 63–84. John Benjamins, Amsterdam-Philadelphia (2015).
- **Gledhill C., Patin S., Zimina M.:** Lexico-grammaire et textométrie : identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. *Corpus* 17 (2017).
- **Fleury, S.**, **Zimina, M.:** Trameur: A Framework for Annotated Text Corpora Exploration. *COLING 2014 the 25th International Conference on Computational Linguistics:* System Demonstrations, August 2016, Dublin, Ireland, pp. 57-61.
- **Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J-P., Obin, N., Pietrandrea, P., Tchobanov**, **A.:** Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014.
- **Lin, Ph. M.S.:** The prosody of formulaic expression in the IBM/Lancaster Spoken English Corpus. *International Journal of Corpus Linguistics* 18 (4), 561–588 (2013).
- **Sitri, F., Tutin, A. (dir.):** Phraséologie et genres de discours. Patrons, motifs, routines. *LIDIL 53* (2016).

Thank You

- maria.zimina@eila.univ-paris-diderot.fr
- nicolas.ballier@univ-paris-diderot.fr