Analyse des seuils de constituance de la phraséologie prosodique dans le corpus ***Rhapsodie*** : *période intonative (IPE) et saillance initiale*

MARIA ZIMINA, NICOLAS BALLIER
EA 3967 CLILLAC-ARP

Plan

- **Phraséologie** et **corpus oraux** (état de l'art)
- Corpus ***Rhapsodie*** : annotation prosodique
- **Liens** entre la **prosodie** et des **phénomènes phraséologiques :**
  - **Analyse textométrique** de **données multiannotées**
  - **Récurrences** observées en début des **périodes intonatives**
  - **Spécificités par genre**
- Premières conclusions
- Perspectives

Phraséologie et prosodie : statut expérimental de la recherche

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 expression 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).

Corpus oraux annotés : **ressources singulières**

Projet ***Rhapsodie*** : corpus de référence en français http://projet-rhapsodie.fr

Laboratoires porteurs :

- MODYCO
- IRCAM
- LATTICE
- ERSS
- LPL

Partenaires

- ATILF
- CORAL
- CRDO Paris (Centre de ressources pour la description de l'oral)
- INRIA Bordeaux, Paris
- SYLED

Base de données ***Rhapsodie***

- Un corpus constitué de 57 échantillons de français parlé
- (5 minutes en moyenne), soit 3 heures de parole (33 000 mots, 89 locuteurs) munies d'une transcription orthographique et phonétique.
- Annotations micro / macro syntaxique et prosodique (plus de 60 couches d'annotations)
- Modélisation de l'interface prosodie, syntaxe, discours en français parlé

Rhapsodie : interface prosodie / syntaxe / discours

- La compréhension du rôle que jouent les indices intonosyntaxiques dans la segmentation du continuum sonore en unités informationnelles et discursives
- La modélisation de l'interface prosodie, syntaxe, discours en français parlé
- La base des données prosodiques alignées sur le temps et des données syntaxiques calibrées sur les tokens syntaxiques.
- …

Annotation prosodique dans *Rhapsodie :* *Exemple* (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

Méthodologie :

Combinaison de l'annotation manuelle (*proéminence* et *disfluences*) et automatique :

1) Annotation manuelle (critères formels acoustiques et perceptifs)
2) Caractérisation automatique des constituants prosodiques à partir de l'annotation manuelle
3) Stylisation automatique de contours mélodiques et l'annotation des tons liés aux constituants prosodiques

Structures prosodiques dans *Rhapsodie*

Intonational Periods (IPE)

Intonational PAckages (IPA)

Rhythmic Groups (RG)

Metrical Feet (MF)

Syllabes (avec la proéminence : non-proéminent : **o**, fort : **S**, faible : **W**

Nature exploratoire de la recherche

- Etablir des **liens** entre la **prosodie** et des phénomènes **phraséologiques**
- Plus de **60 couches** annotations dans ***Rhapsodie***
  - morpho-syntaxiques
  - syntaxiques
  - macro-syntaxiques
  - prosodiques
- Point de départ ?
- Options théoriques, pratiques, outils

Options théoriques (1/2)

- Analyse de séquences préfabriquées qui constituent des **objets phraséologiques plus larges que les expressions figées classiques**

- Mise en relation de ces « prêts à dire » avec les caractéristiques du **genre discursif** dans lequel ces séquences apparaissent

Sitri, F., Tutin, A. (dir.): Phraséologie et genres de discours. Patrons, motifs, routines. *LIDIL 53* (2016).

Options théoriques (2/2)

- *CLILLAC-ARP* : utilisation des corpus pour identifier les **« schémas lexicogrammaticaux »**, qui représentent les unités de structure pour la construction des **discours spécialisés**
- Chaque schéma LG est composé d'un ou plusieurs élément(s) obligatoire(s) (le « **pivot** ») et d'un ou plusieurs éléments variables (le « *paradigme* »).
- Les **imbrications de schémas LG** mènent à la création de chaînes discursives plus étendues. **(Gledhill *et al.*, 2017)**

Exemple :

<*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*>

Données *Rhapsodie*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	TextID	TreeID	TokenID	Token	Lemma	POS	Mode	Tense	Person	Number	Gender	Gov_rection	Type_rection	Gov_para	Type_para	Gov_inher	Type_inher	Gov_junc	Type_jun	Gov_junc-inhe	Type_junc-in
2	M2006	1	1	bonjour	bonjour	B_I						0	root								
3	M2006	1	2																		
4	M2006	1	3	Eric	Eric	B_N				sg	masc	0	root								
5	M2006																				
6	M2006	2	1	bonjour	bonjour	B_N				sg	masc	0	root								
7	M2006	2	2																		
8	M2006	2	3	à	à	B_Pre						1	dep								
9	M2006	2	4																		
10	M2006	2	5	tous	tous	B_Pro			3	pl	masc	3	dep								
11	M2006																				
12	M2006	3	1	nouvelle	nouveau	B_Adj				sg	fem	3	dep								
13	M2006	3	2																		
14	M2006	3	3	nuit	nuit	B_N				sg	fem	0	root								
15	M2006	3	4																		
16	M2006	3	5	de	de	B_Pre						3	dep								
17	M2006	3	6																		
18	M2006	3	7	pillage	pillage	B_N				sg	masc	5	dep								
19	M2006	3	8																		
20	M2006	3	9	et	et	B_J												5	junc		
21	M2006	3	10																		
22	M2006	3	11	d	de	B_Pre								5	para_coord	3	dep_inherited	9	junc		

Ces données sont constituées par un certain nombre de textes (l'identifiant du texte est visible dans la première colonne), chacun d'eux est segmenté en *unités illocutoires* (seconde colonne), chacune d'elle est segmentée en *tokens* (troisième colonne), chacun d'eux est annoté (les autres colonnes)

Base textométrique *Rhapsodie* (S. Fleury)

Transcodage des données : *Trame/Cadre*

```
<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></item>
```

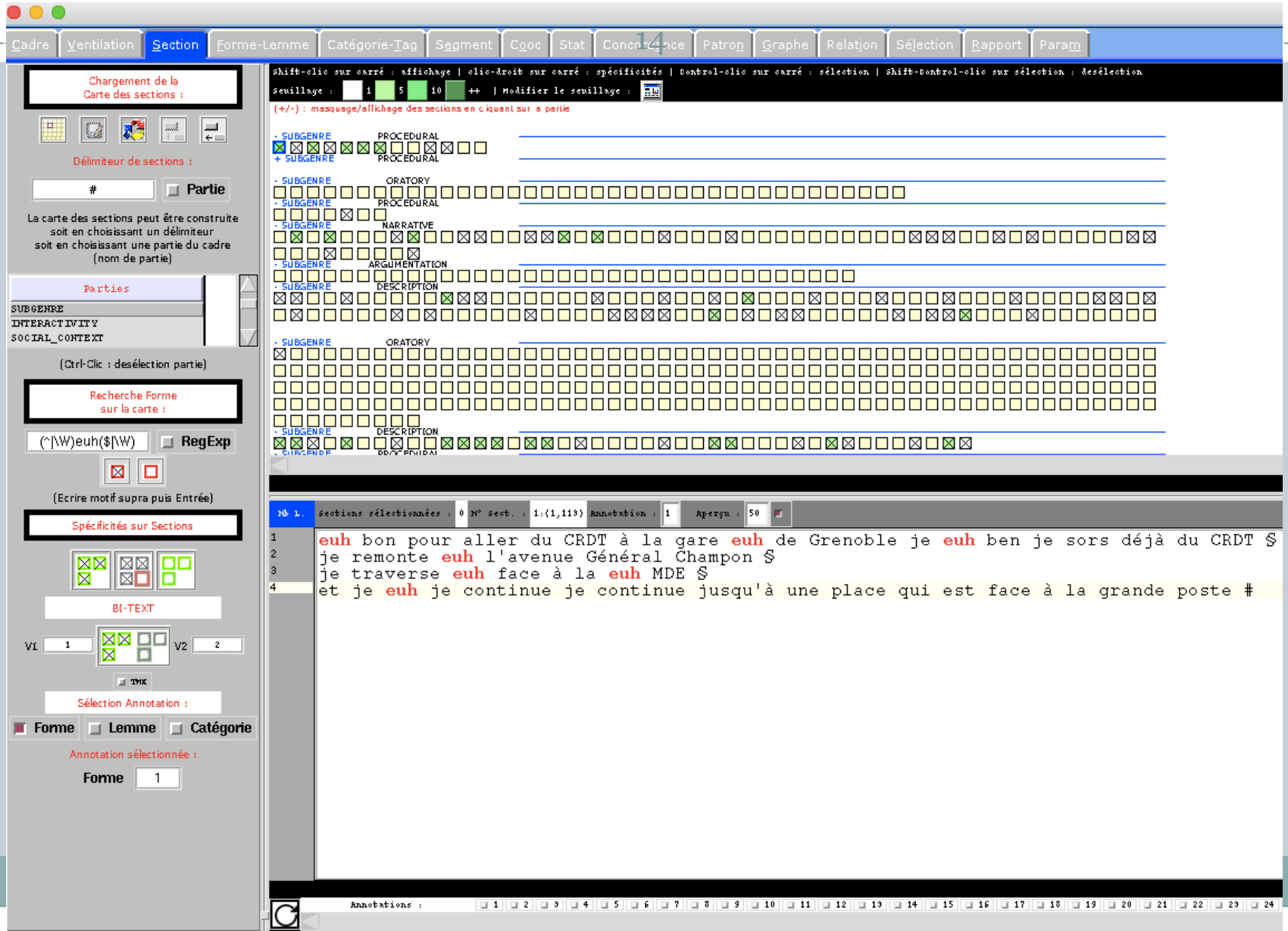
Trame

```
<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"/>
```

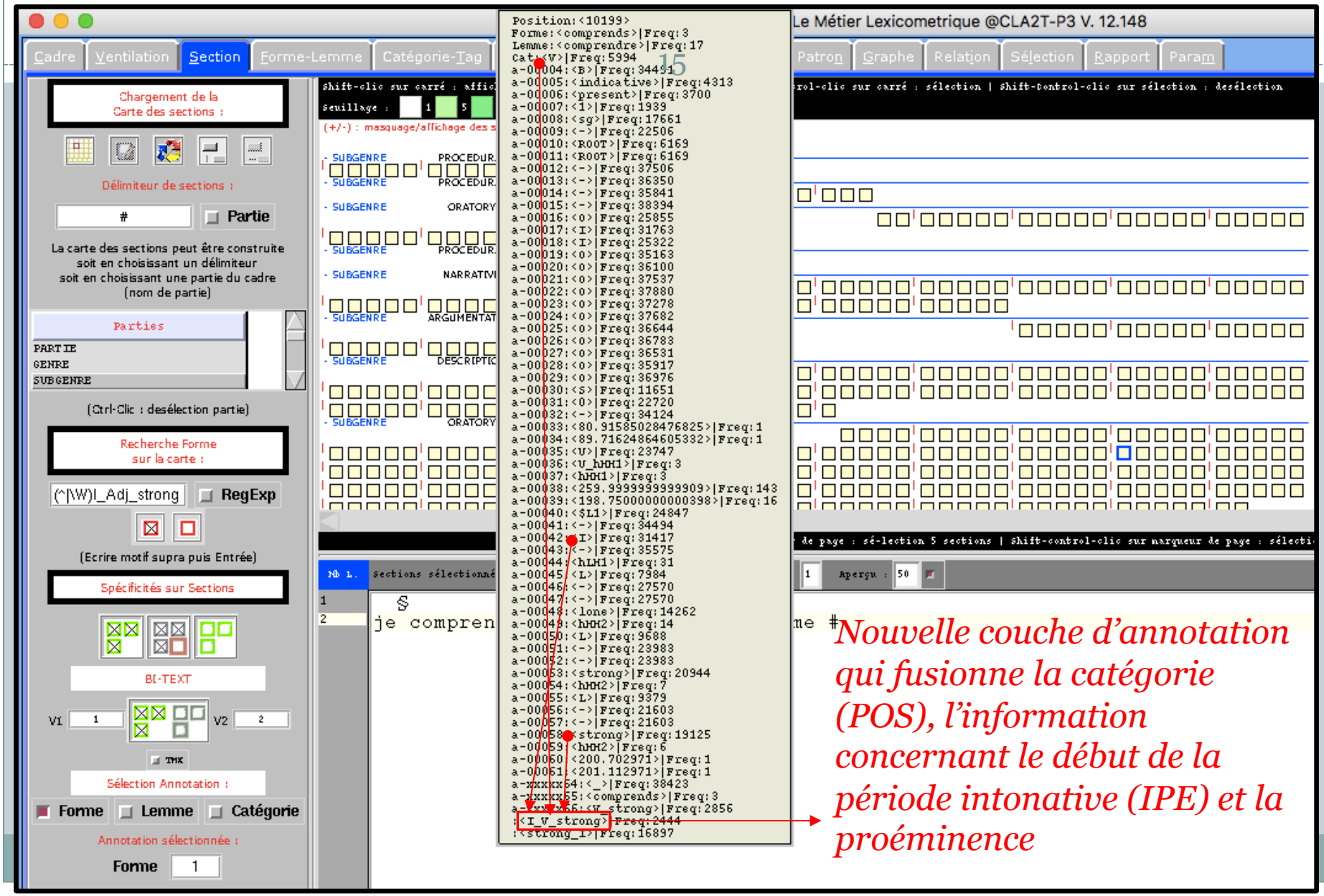
Cadre

Le processus de transcodage des données issues du projet ***Rhapsodie*** (extrait). Dans la dernière version de la base, chaque item de la *Trame* est associé à **61 niveaux d'annotation** (prosodie, micro et macro-syntaxe).

Base ***Rhapsodie*** importée dans ***Le Trameur***



Annotations croisées dans *Le Trameur*



Nouvelle couche d'annotation qui fusionne la catégorie (POS), l'information concernant le début de la période intonative (IPE) et la proéminence

Hypothèses

- La segmentation en **constituants prosodiques** est liée à l'**organisation discursive** (en relation avec le genre).
- La constitution des périodes intonatives (IPEs) a des régularités qui reflètent **la présence des unités de structures** (« schémas LG ») pour la construction de l'oral.
- L'analyse de la **saillance initiale** de l'**IPE** est le point d'entrée.

IPEs : saillance initiale

Catégorie	Saillance initiale (B_IPE)	Total dans le corpus
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 : segments avec saillance initiale

Segment répété (catégories)	Saillance initiale (B_IPE)	Total dans le corpus
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.

La méthode des **spécificités** (Lebart et Salem, 1994. *Statistique textuelle*.)

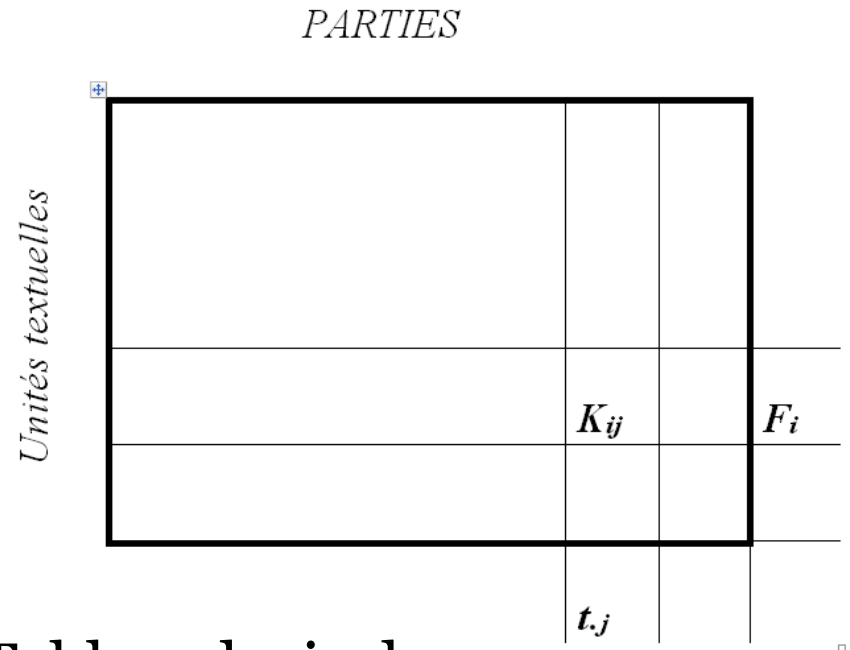
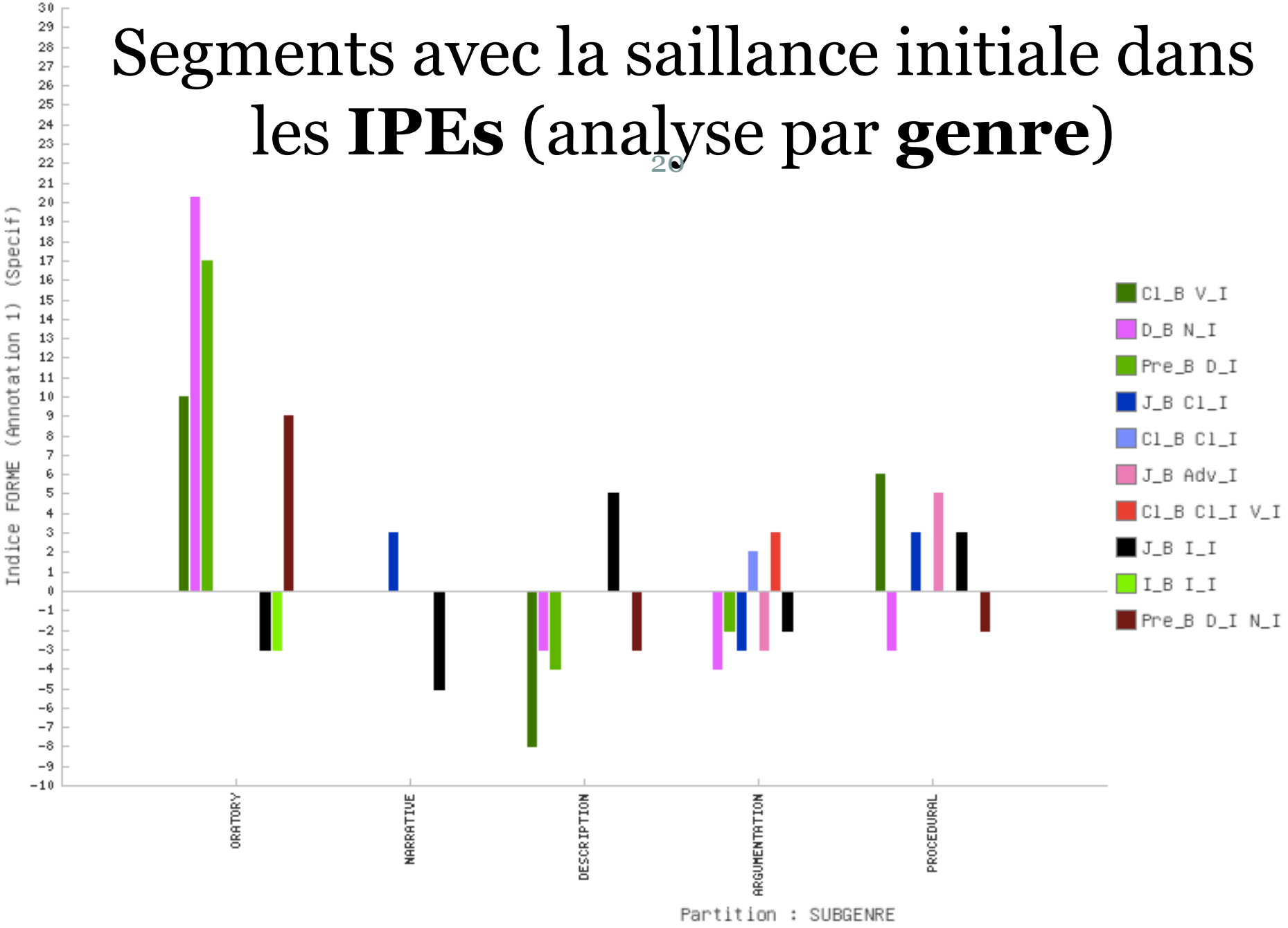


Tableau lexical :

K_{ij} : fréquence de l'unité ***j*** dans la partie ***i***
F_i : fréquence de ***i*** dans le corpus
t_j : taille de la partie ***j***

Si l'effectif K_{ij} ne se situe pas dans les limites de ce que le **calcul probabiliste** permettait d'espérer, on calcule un *indice de spécificité* : *sur-emploi* ou *sous-emploi* de l'unité (spec. ***±xx***)

Segments avec la saillance initiale dans les **IPEs** (analyse par **genre**)



Contextes avec saillance initiale

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 #
énoncés performatifs	*focus*	*instructions*

Le caractère « # » marque le début de la période intonative (**IPE**)

Premières conclusions

- L'analyse textométrique de constituants prosodiques (**IPEs**) a fait émerger des **éléments caractéristiques** dont la **saillance initiale** varie en fonction de **genres discursifs**.

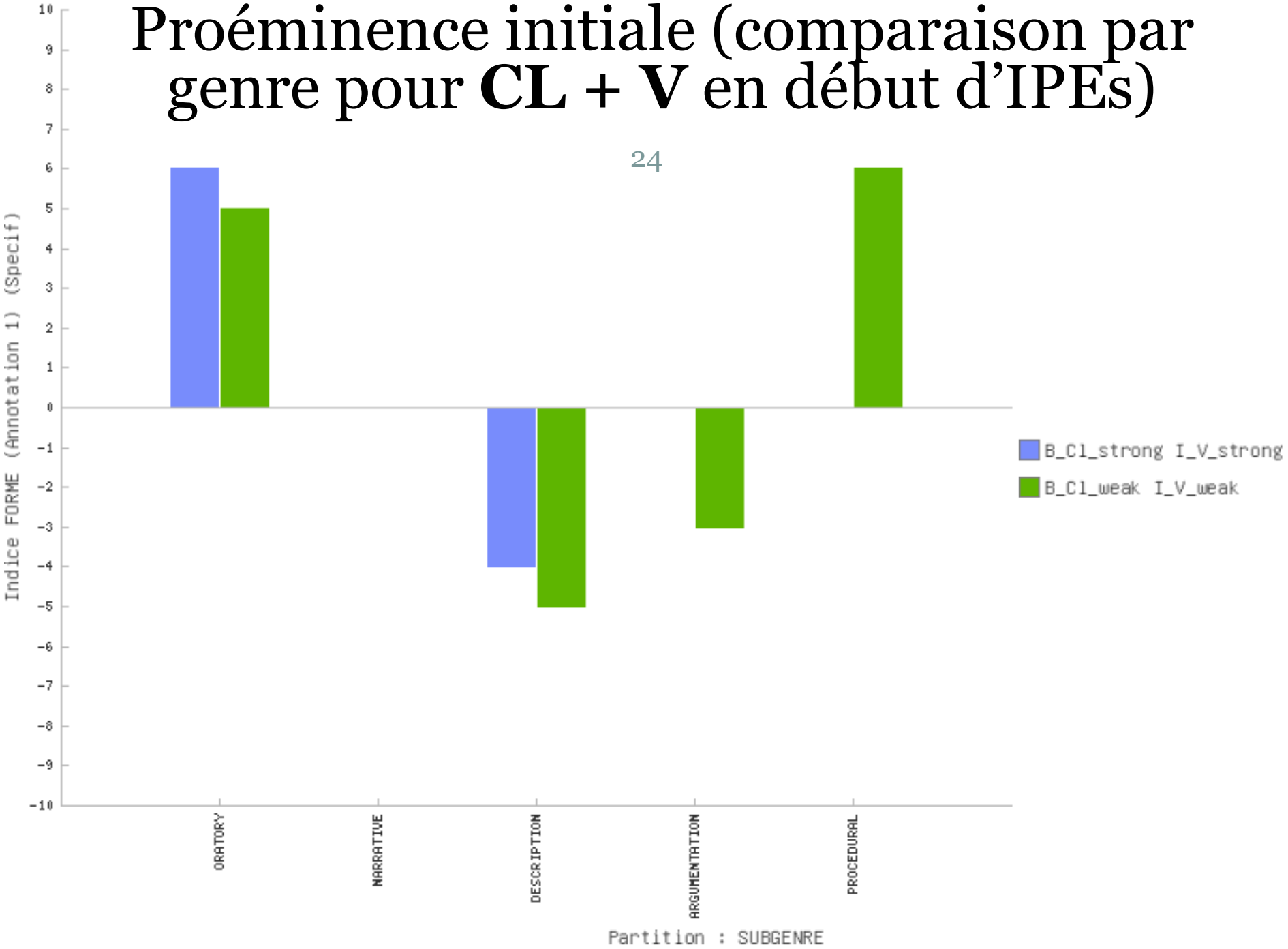
- Les « **pivots** » recensés correspondraient aux éléments stables de **schémas LG** qui structurent l'oral (*"je salue"*, *"elle souhaite"*, *"il faut"*, *"on continue"*, etc.).

- La **saillance initiale** reflète les besoins communicationnels (interaction, tours de parole).

Perspectives

- *Prise en compte de plusieurs niveaux d'annotation prosodiques (proéminences, pauses, tonalité, …)*

Proéminence initiale (comparaison par genre pour **CL + V** en début d'IPEs)



Saillance initiale avec PROÉMINENCE : pivot CL + V (genre : *Oratory*)

proéminence FORTE

quatre-vingt-dix-neuf dans l'épreuve # § je pense aux nombreuses victimes de la tempête # et
endeuillées dont nous partageons la peine # § je pense à nos concitoyens # cruellement touchés dans leur
précarité de ce qui nous semblait acquis # § nous voyons combien # tout peut être parfois remis en
dans l'épreuve # faire parler leur coeur # je voudrais dire # merci à tous les Français #
on maîtrisait les règles et les habitudes # § je comprends ces mouvements de l'âme # § pourtant
engagent # et qui garantissent notre avenir # § nous avons choisi ensemble de faire grandir la France dans
et capable de faire reculer la pauvreté # § ce sera tout le sens du combat de la France
toute sa place # pour dire le droit # le faire respecter # avec autorité # avec justice
bien vivre ensemble # § mes chers compatriotes # je mesure l'honneur et la responsabilité # qui m
d'outre-mer # de l'étranger # je souhaite très chaleureusement # une bonne # &

proéminence FAIBLE

mes chers compatriotes # je voudrais d'abord # exprimer ma sympathie #
'assurer la coordination des moyens du pays # § nous mesurons surtout le prix de l'aide fraternelle #
deux mille # est devenu # contemporain immédiat § je suis sûr que beaucoup d'entre vous # vont
an deux mille cent # § ces progrès # ne prendront tout leur sens # que s'ils bénéficient
être le siècle # de l'éthique # § je sais que bien des tragédies aujourd'hui font douter
l'intérieur de chaque nation # une exigence # se fait entendre # toujours plus forte # pour
ces valeurs # § en les faisant vivre # nous serons plus forts pour aborder les temps qui viennent
qui viennent # § la France change # § elle doit le faire au rythme du monde # §
§ en étant fidèle à son génie propre # elle saura conjuguer le changement et la cohésion sociale #
riche de fièvre de passion # d'enthousiasme # elle continue comme hier # à ouvrir # et
§ plus fraternel # plus volontaire # il aura les couleurs # que nous lui donnerons

Perspectives

- *Critères perceptifs utilisés lors de l'annotation manuelle des constituants prosodiques ?*

Références

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).