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

Maria Zimina, Nicolas Ballier

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

Maria Zimina[1] and Nicolas Ballier[1]

[1] Univ Paris Diderot, Sorbonne Paris Cité, CLILLAC-ARP, EA3967, 75013, Paris, France
mzimina@eila.univ-paris-diderot.fr
nicolas.ballier@univ-paris-diderot.fr

**Abstract.** This paper addresses prosodic aspects of phraseology from the point of view of the 'lexicogrammar' approach in *Rhapsodie*, a richly annotated corpus of spoken French. Textometric methods enable the selection of statistically relevant phenomena both in terms of marked prosodic salience and recurrent lexicogrammatical features. Among the possible prosodic characteristics of phraseology, salient initial prominences of prosodic macro-units are considered in this paper. Within the *Rhapsodie* annotation framework, these macro-units correspond to the largest prosodic constituents, the Intonational PEriods (IPE). We have analysed the IPEs bearing the strongest level of initial prosodic salience. The prosodic properties of the units so delineated are discussed, as well as the discourse type of these phenomena. Some recurrent patterns of oratory and procedural speech genres are described using prosody and Part-of-Speech (POS) annotations. We suggest that the role of initial salience of specific prosodic patterns is to facilitate perception and interaction in human speech and to establish the structure of speakers' turns.

**Keywords:** Lexicogrammar, Prosodic Constituents, Salience, Textometrics

## 1 Introduction

### 1.1 Phraseology and Prosody

Our research examines the notions of phraseology and formulaic language in the process of speech production. The terms "phraseology" and "formulaic language" will be discussed in this paper from the point of view of the 'lexicogrammar' approach [1]. From this perspective, our objects of study are predictable and productive sequences of signs called lexicogrammatical patterns (lexical signs, grammatical constructions). Composed of permanent 'pivotal' signs and a more productive 'paradigm', these patterns may be discontinuous and may or may not be syntactic constituents. In contrast with the lexicological approach to phraseology, which studies phraseological phenomena in terms of a continuum ranging from 'free combinations' to 'fixed phrases', the 'lexicogrammar' approach is particularly suited to identifying extended patterns within a particular register or genre [ibid.]. We aim to explore the ways in which prosodic features may correlate with extended lexical patterns, as well as the extent to

which prosody corresponds to patterns which have a particular register or discourse function.

Recent studies have shown that prosodic features extracted from speech databases can be successfully integrated into the investigation of phraseology [2] and its pedagogical applications [3]. Prosody is key to fluency in speech production and reception [ibid.]. Most transcriptions of prosody indicate specific events in speech: boundary tones, pitch accents, disfluent segments, etc. [4]. These speech events coded in spoken corpora are possible candidates for identifying the prosodic characteristics of formulaic language [5] [6].

In this respect, a quantitative analysis of finely annotated spoken corpora facilitates research on the prosodic aspects of phraseology: *"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."* [2]. Unfortunately, large spoken corpora are rarely distributed with a fine-grained prosodic annotation.

For French, a free reference corpus, the *Rhapsodie* speech database (ANR Rhapsodie 07 Corp-030-01), is now available [7]. This syntactic and prosodic treebank is composed of 57 short samples of spoken French (approximately 5 minutes long), orthographically and phonetically transcribed (approximately 33,000 words). This corpus was 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; planned, spontaneous and semi-spontaneous, etc.) [8]. The resource can be downloaded from www.projet-rhapsodie.fr.

### 1.2 A Summary of the Rhapsodie Methodology for Prosodic Annotation

The *Rhapsodie* project follows a bottom-up approach driven by the data [9]. The transcriptions and the annotations are aligned on the speech signal and *Praat* Textgrids [10] are available online [7]. The prosodic annotation is based on the assumption that, out of the total acoustic signal, only certain perceptual cues selected by the listener are relevant for linguistic communication [9]. Following this assumption, 10 research teams collaborated on the following workflow:[1] (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.

This combination of manual and automated annotations allowed a segmentation of speech into prosodic periods [11], which relies on the initial characterization of two types of speech events retained from the manual annotation: prosodic prominence and disfluencies. For illustration purposes, Fig. 1 gives a summary of this prosodic structure [9]. It is organized around rhythmic and melodic components. The hierarchy of constituents includes:

[1] http://www.projet-rhapsodie.fr/laboratoires, last accessed 2017/09/01.

1. Intonational PEriods (IPE)
2. Intonational PAckages (IPA): sub-constituents internal to periods
3. Rhythmic Groups (RG): sub-constituents internal to intonational packages
4. Metrical Feet (MF): sub-constituents inside rhythmic groups
5. Syllables, with Prominence levels, including: 0 (non-prominent), W (weak) and S (strong).

| 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 | | | | | ynvədet | | | vuzetje | | | nɔr | malmã | | ãtrene | | |
| syllable | kvu | swa | je | dəv | ny | yn | və | det | 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 |

**Fig. 1.** Prosodic structure of the *Rhapsodie* speech database corpus [9]

The speech dataset annotated within this perception-driven prosodic annotation opens up new possibilities for the investigation of phraseology. As the link between the "marked status" as a +phrase/expression/formulaic expression etc. and prosodic constituents is still to be revealed, some of our research questions are of an exploratory nature and more than 60 layers of morpho-syntactic, syntactic, macro-syntactic and prosodic annotation in *Rhapsodie* necessarily open new perspectives for the exploration of the prosodic dimension of phraseology. We have decided to focus on the initial structure of IPE macro-units and on the contribution of recurrent patterns of initial prosodic salience to the perception of formulaic language. Previous research has considered various candidates for specific prosodic contours (sometimes called 'melodic clichés' [12]) in the phraseology of spoken discourse [2] [13]. Our quantitative approach focuses on the recurrent prominences observable after speech breaks.

## 2 Exploring Linguistic Fixedness in French on the Basis of Prosodic Features

### 2.1 Textometric Procedures

The segmentation of the *Rhapsodie* speech data into IPE considers melodic variations in time and silent pauses used, regardless of segmental and syntactic constraints [9]. Following our work on the analysis of lexicogrammatical patterns in written texts [14], our study explores the prosody/lexicogrammar interface. In order to discover regular patterns, we conduct a textometric analysis of repeated POS segments in relation to IPE boundaries. At this stage, the goal is to isolate a set of linguistic regularities associated with what is commonly perceived as a strong prosodic boundary [9]. Computation of **characteristic elements** [14], which comes after the first stage, aims to describe perceived regularities with respect to genre and speaking styles, categorized in *Rhapsodie* as 'subgenres' (see Figure 2 below) [7] [9].

Typically, a **hypergeometric model** [16] is the statistical rationale for the computation of indices signaling characteristic POS or POS repetitions within each of the

**corpus parts** (interactive/non-interactive/semi-interactive speech, dialogue / monologue, subgenre, etc.). The computation adapts classical statistical tests [15] that can detect, within each of the parts of a corpus, which elements are used frequently as well as the ones which tend to be rarely used. As a consequence, characteristic elements discovered in this second stage allow for the investigation of candidate formulae in the breaking of the speech flow across different social contexts. Different variables can be analyzed within this approach using textometric software.

### 2.2 Le Trameur Software

The software used in our study is *Le Trameur* [17]. It allows for the intersecting analysis of multiple speech and text annotation layers in various forms of textometric analysis. For example, more than 60 annotations are used in *Rhapsodie*. They are all displayed and processed in a single graphical user interface [18]. *Le Trameur* can be used to automatically re-annotate the dataset and potentially add new annotation layers. Moreover, the researcher can select and manually correct any occurrence of a given tag of the dataset. Various textometric analyses available using this software have allowed us to compare the specific lexicogrammatical regularities of IPE characteristics in several communicative situations. We used frequent characteristic patterns detected as "starters" of the main prosodic constituents to detect formulaic expressions in different speech contexts. The following section illustrates some of the first findings resulting from our experiments with the *Rhapsodie* corpus.

## 3 Characteristic POS Patterns at the Beginning of the IPE in Several Speech Genres of Rhapsodie

### 3.1 Detection of Salient POS and POS Patterns

**Table 1.** The most frequent POS in the IPE initial position of strongest salience (2, 609 occ.)

| POS list | Strongest initial prosodic salience | Total of the POS (any position) |
|---|---|---|
| 1. Cl (Clitic pronoun) | 511 occ. | 4, 179 occ. |
| 2. J (Coordinating conjunction) | 443 occ. | 1, 142 occ. |
| 3. I (Interjection) | 439 occ. | 1, 984 occ. |
| 4. Adv (Adverb) | 287 occ. | 2, 789 occ. |
| 5. Pre (Preposition) | 238 occ. | 3, 443 occ. |
| 6. D (Determiner) | 209 occ. | 4, 080 occ. |
| 7. V (Verb) | 112 occ. | 5, 994 occ. |
| 8. Qu (Relative pronoun) | 97 occ. | 799 occ. |
| 9. CS (Subordinating conjunction) | 74 occ. | 729 occ. |
| 10. N (Noun) | 65 occ. | 6, 317 occ. |

In decreasing order of frequency, clitic pronouns, coordinating conjunctions, interjections, adverbs, prepositions, determiners, verbs, relative pronouns, subordinating conjunctions and nouns are the most frequent POS categories (resulting from morpho-syntactic tagging with *SEM* [19]) that occur at the beginning of the IPE (Freq>50), see Table 1 above.

**Repeated segments computation** [20] is further used to study the specific attractions of these POS categories at the beginning of the IPE contexts. Table 2 presents the most frequent POS recurrences (Freq>50) in the IPE initial position.

**Table 2.** The most frequent POS recurrences (POS N-grams) in the IPE initial position

| POS repeated segment N-gram list | Strongest initial prosodic salience | Total of the N-gram (any position) |
|---|---|---|
| 1. CL + V | 257 occ. | 2, 223 occ. |
| 2. D + N | 129 occ. | 2, 919 occ. |
| 3. Pre + D | 90 occ. | 1, 112 occ. |
| 4. J + Cl | 77 occ. | 164 occ. |
| 5. Cl + Cl | 76 occ. | 525 occ. |
| 6. J + Adv | 70 occ. | 150 occ. |
| 7. Cl + Cl + V | 69 occ. | 479 occ. |
| 8. J + I | 67 occ. | 107 occ. |
| 9. I + I | 60 occ. | 258 occ. |
| 10. Pre + D + N | 55 occ. | 939 occ. |

### 3.2 Characteristic Elements in Different Speech Contexts

**Characteristic elements** [15] describe parts of the corpus displaying these POS repetitions that are significantly more salient or a great deal less salient in a given part of the corpus than in the overall corpus. For example, it is clear from the graphs on Fig. 2 that Cl + V is a positive characteristic element at the beginning of the IPE in the speech contexts of oratory genre (specificity indice: +10). The following examples reveal some lexicogrammatical realizations of this productive pattern in *Rhapsodie* (categories corresponding to **CL + V** are in bold in the following examples):

- Oratory genre: IPE starting with **Cl + V** (rhetorical function: performatives)

– # **je suis** heureux de me retrouver ce soir #
– # **elle salue** la loyauté #
– # **il faut** les faire grandir #
– # **je souhaite** que l'Europe #

Another strong characteristic pattern of this genre is D + N (specificity: +26):

- Oratory genre: IPE starting with **D + N** (rhetorical function: theme-selection)

— # **la démocratie** politique et sociale #
— # **la France** sera ce que nous voudrons qu'elle soit # une nation unie #
— # **le droit** de grève # le droit à l'instruction #
— # **un moment** fort #
— # **l'exigence** de solidarité #

Cl + V is also revealed as a positive characteristic element of initial prosodic salience (specificity: +6) in the procedural speech contexts:

- Procedural genre: IPE starting with **Cl + V** (rhetorical function: instructions)

— # **on passe** devant le le kiosque à journaux #
— # **tu vas** tout droit #
— # **vous continuez** # vous prenez le rond-point tout droit #
— # **on traverse** la rue #
— # **tu descends** toute la pente #

Coordinating conjunction (J) is a characteristic element of initial salience in procedural speech. It is present in three characteristic patterns: J + I (specificity: +5), J + CL (specificity: +3), J + Adv (specificity: +3):

- Procedural genre: IPE starting with **J + I /CL/Adv** (instructions, hesitation)

— # **et vous** allez toujours tout droit #
— # **et vous** suivez toujours la ligne du tram #
— # **et euh** donc je vais jusqu'au & jusqu'à la place Victor Hugo #
— # **et là** je me retrouve euh en effet euh s~ près des rails du tram #
— # **et euh** et ben voilà j'arrive au niveau de la grande place de la gare où … #

## 4 The Role of Formulaic Expressions in the Organization of Speech

Our findings can be summed up as a set of observational statements:

1. Speech boundaries of intonational periods (IPE) can be easily related to the characteristic repetitions of lexico-grammatical patterns revealed by POS recurrences at the beginning of the IPE. This prosodic salience of recurrent initial left-aligned patterns is a valuable property which can be used to explore how prosody is related to formulaic language. For example, in the French reference corpus, intonational periods tend to begin with specific POS categories, the most frequents ones being clitic pronouns, coordinating conjunctions, interjections, adverbs, prepositions, determiners and verbs. These results are to be compared with other available speech corpora data in French.

2. The initial IPEs vary in different social contexts and reflect specific communicative needs of the speakers. However, the pivotal elements of these productive patterns, such as the expression of predicates CL+V, have stable lexico-grammatical realizations in the *Rhapsodie* speech dataset (*"je salue"*, *"elle souhaite"*, *"il faut"*, *"on continue"*, etc.). These elements are regularly reproduced in specific linguistic contexts and reflect regular rhetorical units (performatives, theme-selection, instructions etc.) with predictable/definable discourse functions.
3. The linguistic characteristics of these salient contexts and their degree of semantic unity are quite different. For example, the illocutionary unit § ***# on passe devant le le kiosque à journaux #*** shows a possible extension of the pivotal element *"on passe"* in a specific communicative situation, while # *un moment fort* # has a greater semantic unity. More experiments with the *Rhapsodie* dataset are necessary to explore the precise nature of these phenomena using other annotation levels.

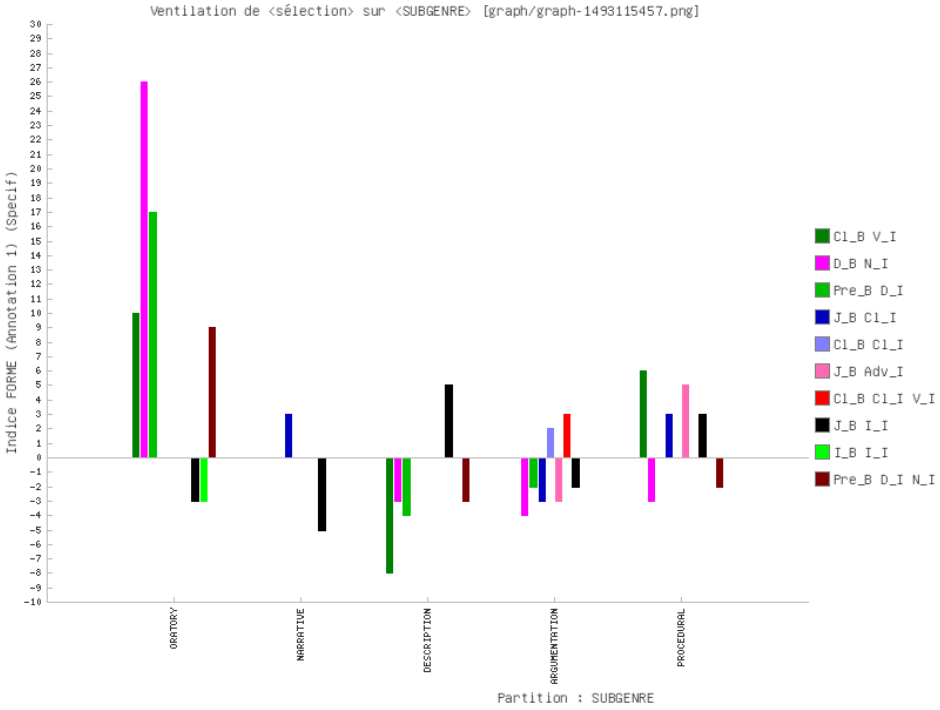


**Fig. 2.** Specificity of POS N-grams at the beginning of the IPE (with positional properties labeled as _B: Beginning, _I: Inside) in several speech genres of *Rhapsodie*

## 5 Future Work

One of the interesting outputs of this study is that it paves the way for a new investigation on inter-annotator agreement in prosodic processing. As manual annotation is

part of prosodic segmentation in *Rhapsodie*, future experiments could consider constraints of memory [21] and processing capabilities of the annotators in determining IPE boundaries when it comes to formulaic language. However, we are convinced that for the present study of the phraseology/prosody interactions, the specificities of the *Rhapsodie* speech database are valuable inputs.

## 6 Conclusion

Prosodic segmentation into intonational periods offers new insights for the observation of the functions of formulaic expressions in speech. A possible place to start with is the identification of annotated IPE boundaries and their correspondence with frequent morpho-syntactic repetitions. In this respect, characteristic POS repetitions at the beginning of intonation periods are more than simply recurrent groups of linguistic units. In our opinion, they represent an observational tool that can be used to investigate how prosodic variations depending upon several factors (interactional need, social context, genres, etc.) are related to formulaic language.

The experiments proposed in our study represent an attempt to account for the uses of these specific patterns after prosodic breaks where the speakers are likely to rely upon the formulaic language for specific communication purposes. A possible interpretation of these prosodic features naturally comes from the analysis of different speech contexts where the initial salience of specific prosodic phraseology facilitates speech perception. In this respect, recurrent patterns are likely to reflect strong speech signals to which speakers and listeners respond in a distinct way, showing an important influence of intrinsic experience of language acquisition in the structuring of speaker and listener interaction, speakers' turns, etc. This type of analysis can be further extended to include all the other prosodic characteristics (tone, pause length, etc.) available in the *Rhapsodie* speech dataset.

## References

1. Gledhill, C.: The 'lexicogrammar' approach to analysing phraseology and collocation in ESP texts. ASp (Anglais de Spécialité) 59, 05–23 (2011).
2. Lin, Ph. M.S.: 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 (2013).
3. 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).
4. Yoo, H-Y, Delais-Roussarie, E. (eds.): Actes d'IDP 2009, Paris, France, September (2009), http://makino.linguist.jussieu.fr/idp09/actes_fr.html, last accessed 2017/09/01.
5. Granger, S.: 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 (2005).

6. Wray, A.: Formulaic language: Pushing the boundaries, Oxford University Press, Oxford (2008).
7. RHAPSODIE Homepage, http://www.projet-rhapsodie.fr/, last accessed 2017/09/01.
8. Lacheret, A., Kahane, S., Pietrandrea, P. (eds.): Rhapsodie: a prosodic and syntactic treebank of spoken French, John Benjamins, Amsterdam-Philadelphia (2017).
9. 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, http://www.lrec-conf.org/proceedings/lrec2014/index.html, last accessed 2017/09/01.
10. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program]. Version 6.0.29, retrieved 2017/09/01 from http://www.praat.org/
11. Lacheret, A., Victorri, B.: La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. Verbum 24 (1-2), 55–73 (2002).
12. Kawaguchi, Y., Fonagy, I., Moriguchi, T.: Prosody and Syntax: Cross-linguistic Perspectives, John Benjamins, Amsterdam-Philadelphia (2006).
13. Cheng, W., Greaves, C., Warren, M.: A corpus-driven study of discourse intonation: the Hong Kong corpus of spoken English (prosodic). John Benjamins: Amsterdam-Philadelphia (2008).
14. Gledhill C., Patin S., Zimina M.: Identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. CORPUS 17 (2017), https://corpus.revues.org/, last accessed 2017/09/01.
15. Lebart, L., Salem, A., Berry, L.: Exploring Textual Data. Kluwer Academic Publishers, Dordrecht, Boston (1998).
16. Lebart, L., Salem, A., Berry, L.: Recent developments in the statistical processing of textual data. Applied Stochastic Models and Data Analysis 7 (1), 47–62 (1991).
17. LE TRAMEUR Homepage, http://www.tal.univ-paris3.fr/trameur/, last accessed 2017/09/01.
18. Fleury, S., Zimina, M.: Trameur: A Framework for Annotated Text Corpora Exploration. In: Proceedings of 25th International Conference on Computational Linguistics (COLING 2014), August 2014, Dublin, Ireland. Proceedings of COLING 2014 the 25th International Conference on Computational Linguistics: System Demonstrations, August 2014, Dublin, Ireland, pp.57–61 (2014), http://www.aclweb.org/anthology/C14-2013.pdf, last accessed 2017/09/01.
19. SEM Homepage, http://www.lattice.cnrs.fr/sites/itellier/SEM.html, last accessed 2017/09/01.
20. Salem, A.: Pratique des segments répétés. Essai de statistique textuelle, Klincksieck, Paris (1987).
21. Gurevich, O., Johnson, M. A., Goldberg, A. E.: Incidental verbatim memory for language. Language and Cognition 2 (1), 45–78 (2010).