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

Maria Zimina-Poirot, Nicolas Ballier

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

Maria Zimina[1], Nicolas Ballier[2]

[1]Université Paris Diderot – mzimina@eila.univ-paris-diderot.fr

[2]Université Paris Diderot – nicolas.ballier@univ-paris-diderot.fr

Abstract

This paper focuses on specific quantitative characteristics of spoken language phraseology in the *Rhapsodie* speech database (ANR *Rhapsodie* 07 Corp-030-01). A recent study (Zimina & Ballier, 2017) has shown that prosodic segmentation into IPE: Intonational PEriods (segments of speech with distinctive pitch and rhythm contours) available within the *Rhapsodie* database offers new insights for the observation of the functions of formulaic expressions in speech. Recurrent lexicogrammatical patterns at the beginning of Intonational PEriods (IPE) are strongly related to spoken formulaic language. These variations of initial salience depend upon several factors (interactional needs, social context, genres, etc.). Further experiments have shown that initially salient patterns also have specific prosodic characteristics in terms of prominence (prosodic stress) across major speech genres of the *Rhapsodie* dataset (oratory, narrative, description, argumentation, procedural) and corresponding speaking tasks. These specific prosodic characteristics are likely to reflect communicative needs of speakers and listeners (interactions, uptakes, speaking turns, etc.).

Keywords: phraseology, prosodic constituents, prominence, salience, textometrics

1. Introduction

Our research examines the notions of phraseology and formulaic language in speech production on the basis of prosodic transcriptions indicating specific events in speech: boundary tones, pitch accents, disfluent segments, etc. (Yoo et Delais-Roussarie, 2009). We believe that such speech events coded in spoken corpora are relevant for identifying the prosodic characteristics of formulaic language.

Corpus-based studies of phraseology often exploit recurrent patterns detected using repeated segments, co-occurrences and pattern-matching techniques to explore formulaic strings of written texts (Granger, 2005; Sitri et Tutin, 2016). This approach seems equally applicable to oral discourse. Following this approach, our initial objects of study are predictable and productive sequences of signs called *lexicogrammatical patterns* (lexical signs, grammatical constructions). Made of permanent **'pivotal'** signs and a more productive '*paradigm*', these patterns may be discontinuous and may or may not be syntactic constituents (Gledhill, 2011; Gledhill et al., 2017). For example:

§ 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 …
§ mais bien sûr donc <**c'est pour** *ça* bien sûr bien sûr **que** *je parlais* oui XXX …
§ **c'est pour** *cela* **que** *je tenais* à vous rencontrer la veille de notre fête …

We then 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. These lexicogrammatical patterns combined with

prosodic features extracted from speech databases are possible methodological tools for identifying phraseological characteristics of oral discourse.

2. The *Rhapsodie* speech database

Large spoken corpora are rarely distributed with a fine-grained prosodic annotation. Fortunately, for French, a reference corpus, the *Rhapsodie* speech database (ANR Rhapsodie 07 Corp-030-01), is freely available online (http://www.projet-rhapsodie.fr/). 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).

2.1. Database structure

The *Rhapsodie* data file is available online in tabular form. The corpus data is structured as follows: all *Rhapsodie* texts are first identified by codes. Each text is further divided into separate units and segmented into tokens. The remaining columns display more than 60 linguistic annotations of the tokens, including microsyntax (rection, dependency, constituency), macrosyntax and prosody. This data set was transformed from the spreadsheet format into a *Trameur* base file using regular expressions (Fleury, 2013). The data structure is composed of two parts: (1) a *Thread*, which is a list of items with position identifiers; (2) a *Frame*, which is a list of corpus partitions defined on the *Thread*. Each partition has a name and a list of named constituents identified through their first and last token positions on the *Thread*. Thus, each annotated token from the *Rhapsodie* corpus becomes an item identified by its position on the *Thread* (Fleury et Zimina, 2014).

2.2. Prosodic annotation

The corpus covers several discourse types and speaking styles: oratory, narrative, description, argumentation, procedural; interactive, semi-interactive and non-interactive; public and private, planned, spontaneous and semi-spontaneous, etc. (Lacheret et al., 2017). The transcriptions and the annotations are aligned on the speech signal (Lacheret et al., 2014). A combination of manual and automated annotations allowed a segmentation of speech into prosodic periods (Lacheret et Victorri, 2002), which relies on the initial characterization of two types of speech events retained from the manual annotation: prosodic prominence and disfluencies.

Organized around rhythmic and melodic components, the hierarchy of prosodic constituents includes: Intonational PEriods (IPE); Intonational PAckages (IPA): sub-constituents internal to periods; Rhythmic Groups (RG): sub-constituents internal to intonational packages; Metrical Feet (MF): sub-constituents inside rhythmic groups; Syllables, with Prominence levels, including: 0 (non-prominent), W (weak) and S (strong).

3. Quantitative analysis of the prosodic dimension 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.

3.1. Preliminary research

Previous research on spoken phraseological units (Lin, 2013) did not take into account the prosodic hierarchy, in other words, the various sizes of the prosodic constituents (Nespor et

Vogel, 2007). We first replicated this methodology, which consisted in describing the prosodic characterisations of phraseological units attested in speech-to-text transcription. For example, in the categorization of the stress hierarchy, we can distinguish between stressed (*strong*) and less prominent (*weak*) syllables.

Preliminary analyses of repeated segments from the *Rhapsodie* corpus, such as *jeune fille* (F=20) or *je veux dire* (F=21) led us to observe the non-congruency of the recurrence of prosodic features, such as prominence, and traditional phraseological units recovered from transcribed speech data:

… une|*dis-weak* **jeune|*strong*** **fille|*weak*** euh|*weak* habillée|*weak* tout|*strong* en|*strong* …
… c|*tail*'|*tail* est|*tail* une|*tail* **jeune|*tail*** **fille|*tail*** # pauvre|*strong* et|*strong* affamée|*strong* …
… dans|*strong* la|*strong* rue|*strong* avec|*weak* une|*weak* **jeune|*weak*** **fille|*filled-dis*** #
§ vous|*weak* voyez|*weak* ce|*weak* que|*strong* **je|*strong*** **veux|*strong*** **dire|*strong*** #
§ euh|*tail* vous|*weak* voyez|*weak* ce|*weak* que|*weak* **je|*weak*** **veux|*weak*** **dire|*weak*** §

Our first analyses of these examples made us simultaneously consider multiple prosodic properties of these collocations, as well as the necessity of taking into account the hierarchy of prosodic units corresponding to these speech contexts.

3.2. Prosodic constituents: a genre-based analysis of lexicogrammatical recurrence and initial salience

A recent study (Zimina et Ballier, 2017) has shown that prosodic segmentation into IPE: Intonational PEriods (segments of speech with distinctive pitch and rhythm contours) available within the *Rhapsodie* database offers new insights for the observation of the functions of formulaic expressions in speech. Lexicogrammatical patterns at the beginning of IPE are strongly related to spoken formulaic language. Recurrent prominences observable after speech breaks can be revealed by textometric analysis of **repeated Part-Of-Speech (POS) segments** (Salem, 1987) at the beginning of the IPEs. This method can be used to isolate a set of pivotal elements associated with what is commonly perceived as a strong prosodic boundary. Computation of **characteristic elements** (Lebart et al., 1998), applied to **repeated segments** (Salem, 1987), describes these regularities with respect to genre and speaking styles, categorized in *Rhapsodie* as 'subgenres' (Lacheret et al., 2014).

Discovered variations of initial salience depend upon several factors (interactional needs, social context, genres, etc.). They reflect specific communicative needs of the speakers. For example, **Cl + V** is a positive characteristic element at the beginning of the IPE in the speech contexts of oratory genre (specificity index: +10). The following examples reveal some lexicogrammatical realizations of this productive pattern in *Rhapsodie* (categories corresponding to **Cl + V** appear in bold):

il faut les faire grandir
§ **ce sera** un coup franc
je souhaite que l'Europe

These pivotal elements reflect the structure of regular rhetorical units with a predictable/definable discourse function (performative utterances).

3.4. Prosodic salience and prominence: systemic combination

To fine-tune our study of regular prosodic features of phraseology, we have added another layer of analysis, namely the prominence of final syllables. For these purposes, we have combined three annotation levels: (1) the positional properties of units within the IPE structure with *BILOU* tags (*Beginning*, *Inside*, *Last*, *Unit*-length and *Outside*); (2) POS tags;

(3) final stress prominence: *strong*, *weak*, *pause_* and *%* (inaudible or non-transcribed due to overlap). The base file has been automatically re-annotated with *Le Trameur* to add this new combined annotation layer to the *Rhapsodie* data. Repeated segments have then revealed that at the beginning (B) of IPEs, the final prominence of the pivot **Cl + V** is either *weak weak* (F=114) or *strong strong* (F=95). On Figure 1, the results of characteristic elements analysis in different speech genres show that these weak realisations of final syllables are positive characteristic elements (specificity index: +06) of procedural utterances, such as instructions in travel planning, while both strong (+06) and weak (+05) realisations of **Cl + V** are characteristic elements of initial prosodic salience in oratory genre.

A finer-grained analysis according to speaking task, presented on Figure 2, shows that this prosodic richness can be attributed to the subgenre of political discourse, a well-known subtle and complex discourse phenomenon (Dorna, 1995; Mayaffre, 2002). For 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 **Cl + V** also corresponds to emphatic realisations at the beginning of IPE in sermons (+03), as evidenced in Figure 2. Similarly, advertising favours recurrent overuse of stressed syllables in this position (+04).

4. Conclusions

The prosodic hierarchy (Nespor et Vogel, 2007) acknowledges several layers of granularity, from the prosodic utterance to the phoneme. For our investigation, the IPE was the best initial candidate for the proper level of granularity in the prosodic hierarchy. There were structural reasons for this, such as the fact that speaking turns were likely to be signalled in initial position of IPE. The textometric analysis of this prosodic constituent of the prosodic hierarchy has shown revealing features such as the limited distribution of POS categories in the initial position, as well as the role of prosodic prominence (stressed syllables) and its relevance for the distinction of speech genres. It appears that the 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. In all likelihood, the initial characteristic distributions with specific prosodic characteristics correspond to communicative needs (interactions, uptakes, speaking turns, etc.).

Because of the complexity of the *Rhapsodie* speech database, we regard these explorations as preliminary: we have only based our analysis on few layers of annotations. Besides, other layers of granularity within the prosodic hierarchy might also be relevant: Intonational PAckage (IPA), Rhythmic Group (RG), etc. We also surmise that other variables (channel, planning type, event structure: monologal vs. dialogal tasks) are likely to reveal related features.

5. Future research

In future work, various lines of investigation can be pursued, such as the examination of the various layers of the prosodic hierarchy. Looking for collocational structures may lead us to question the recurrence of patterns within prosodic units, in other words, the embedding of prosodic constituents or the complexity of the boundaries of the constituents across the layers of the prosodic hierarchy. Other correlates might be considered such as duration and prosodic

contours (the whole corpus also includes the sound files). It may well be the case that annotators were influenced by the genre of the recordings, and auditory analysis across genres based on our characteristic elements analysis may nuance the levels of prominence assigned by the annotators. If the identification of these collocational prosodic patterns is robust, it should also remain transparent and decisive for subjects when resynthesized. The acoustic signal can be modified so as to erase lexical contents, only keeping the melody (humming). Resorting to humming should enable us to test the relevance of prosodic sequences which should robustly remain identifiable as characteristic signals of collocations in perception tests.

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

Fleury, S., Zimina, M. (2014). Trameur: A Framework for Annotated Text Corpora Exploration. In: *Proceedings of 25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, Ireland, pp.57–61, http://www.aclweb.org/anthology/C14-2013.pdf

Gledhill, C. (2011). The 'lexicogrammar' approach to analysing phraseology and collocation in ESP texts. *ASp (Anglais de Spécialité)* 59, 05–23.

Gledhill C., Patin S., Zimina M. (2017). Identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. *CORPUS* 17, 113–143.

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.

Lacheret, A., Kahane, S., Pietrandrea, P. (eds.) (2017). *Rhapsodie: a prosodic and syntactic treebank of spoken French*, John Benjamins, Amsterdam-Philadelphia.

Lacheret, A., Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum* 24 (1-2), 55–73.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht, Boston.

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.

Mayaffre, D. (2002). Discours politique, genres et individuation socio-linguistique. In: Morin, A., Sébilot, P. (eds.). *Actes des JADT 2002*, Saint-Malo, France, IRISA-INRIA, pp.517–529.

Nespor, M., Vogel, I. (2007). *Prosodic Phonology*. Berlin. Mouton De Gruyter.

RHAPSODIE Homepage, http://www.projet-rhapsodie.fr

Salem, A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Klincksieck, Paris.

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

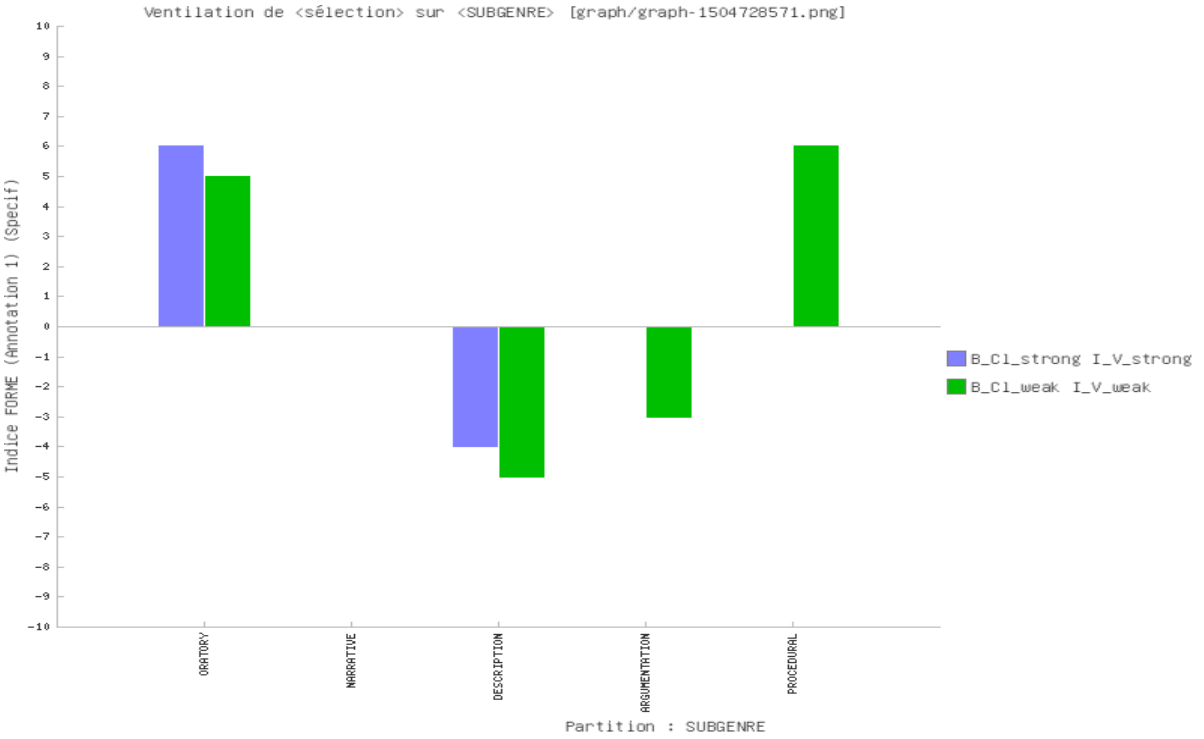

Figure 1: Characteristic elements of initial prominence across speech genres: Clit + Verb in initial position of Intonational PEriods (IPE)

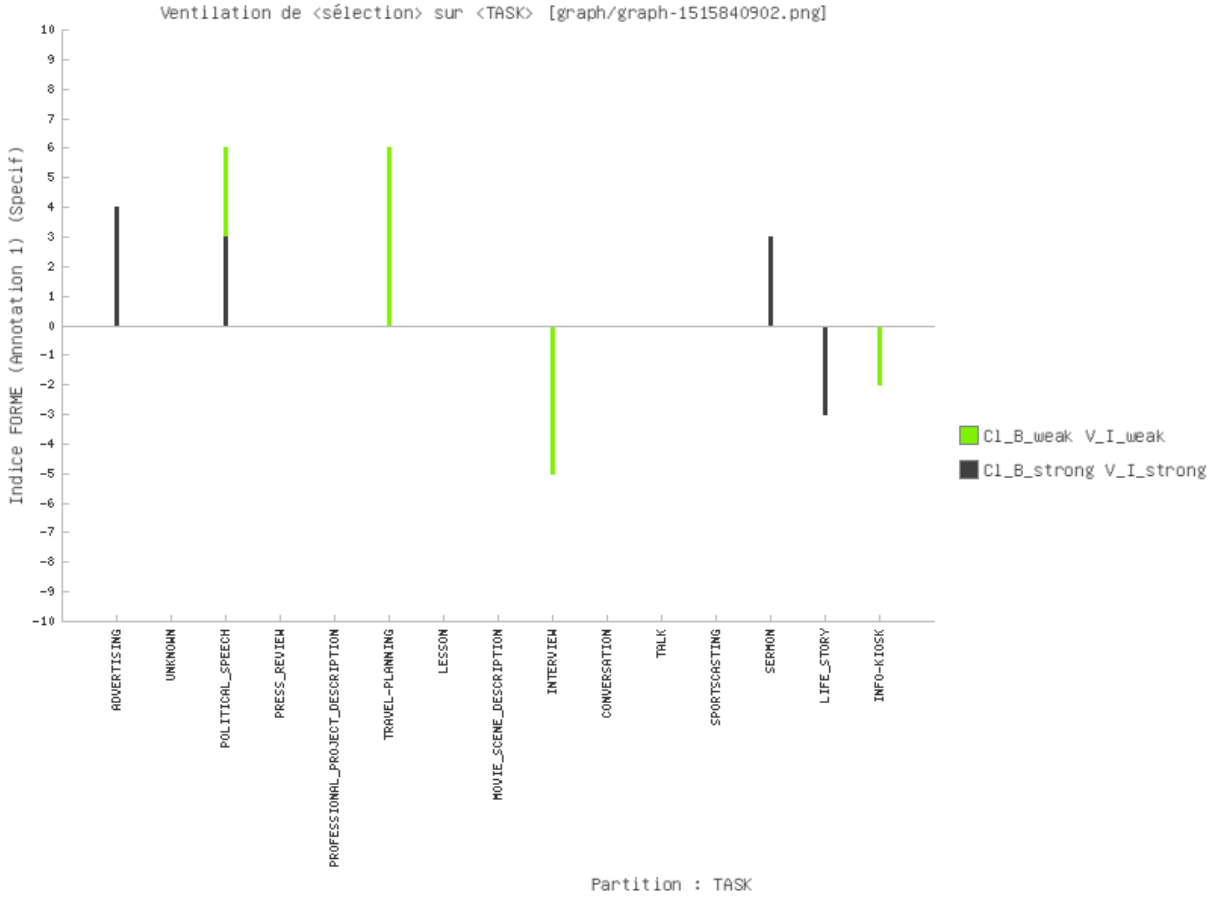

Figure 2: Characteristic elements of initial prominence in different speaking tasks: Clit + Verb in initial position of Intonational PEriods (IPE)