# Fonctions avancées de l'analyse textométrique pour les corpus multi-annotés

*Maria Zimina-Poirot, CLILLAC-ARP, Paris Diderot*

***ITrameur***

**Description générale :**
Cette session de travail permet d'explorer un corpus multi-annoté de français parlé *RHAPSODIE* (Lacheret et al. 2014).

Ce corpus de référence en français parlé disponible librement (http://projet-rhapsodie.fr) permet de :
1) comprendre le rôle que jouent les indices intonosyntaxiques dans la segmentation du continuum sonore en unités informationnelles et discursives ;
2) modéliser l'interface prosodie/syntaxe/discours en français parlé ;
3) mener des analyses fines des données prosodiques alignées sur le temps, et des données syntaxiques calibrées sur les tokens syntaxiques.

Nous montrons, par exemple, comment les données de *RHAPSODIE* permettent d'étudier des *séquences préfabriquées* (« je salue », « il faut », « on passe »), qui constituent des objets phraséologiques plus larges que les expressions figées classiques, en les mettant en relation avec le genre discursif (énoncés performatifs, instructions, etc.).

**Corpus étudié :**
TREEBANK *RHAPSODIE* : http://projet-rhapsodie.fr
- 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.
- Chaque échantillon est catégorisé suivant différents points de vue : genre, type de conversation, etc.
- Annotations micro/macro syntaxique et prosodique (plus de 60 couches d'annotation).
- Modélisation de l'interface prosodie, syntaxe, discours en français parlé.

- Base textométrique *RHAPSODIE* avec 61 couches d'annotations (S. Fleury) : http://www.tal.univ-paris3.fr/trameur/iTrameur/itrameur-BASE-RHAPSODIE-full.txt

**Ressources :**
Base(s) textométrique(s) disponible(s) pour l'atelier sur le site :
*iTrameur* : http://www.tal.univ-paris3.fr/trameur/iTrameur/
Description (Serge Fleury) :
http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur-v8.pdf
http://www.tal.univ-paris3.fr/trameur/corli-trameur/Le-Trameur.pdf (pages 6-16, 37-42)

**Contexte de l'étude :**
A partir de l'analyse du corpus de français parlé *RHAPSODIE* (Lacheret et al. 2014), nous voudrions proposer quelques points d'entrée dans l'analyse des structures phraséologiques du point de vue de leur marquage prosodique en considérant la macro-unité la plus grande au sein de la hiérarchie prosodique du corpus : la période intonative (Zimina et Ballier, 2017 ; 2018). Nous mettrons à contribution la richesse d'annotation du corpus et la spécificité du modèle BILOU, qui assigne une position aux marqueurs au sein des unités (initiale : B) mais permet également d'analyser la labilité du rapport à la frontière des constituants (à l'extérieur de l'unité : O).

**Exploration 1 (se familiariser avec le système d'annotation et les variables d'analyse : partitions) :**
- Afficher le système de partitions du corpus *RHAPSODIE*. Quelle est l'unité minimale du *Cadre ?*
- Quel niveau d'annotation est réservé à l'étiquetage morpho-syntaxique ?

Nous considérons 13 parties du discours (https://www.projet-rhapsodie.fr/wp-content/uploads/2017/04/Protocole-de-codage-microsyntaxique-2013-10-1.pdf) :
-V pour les verbes
-N pour les noms
-Adj pour les adjectifs
-Adv pour les adverbes
-Pre pour les prépositions
-CS pour les conjonctions de subordination
-J pour les joncteurs: conjonctions de coordinations et d'autres éléments de liaison
-D pour les déterminants
-I pour les interjections, y compris des marqueurs de discours
-Qu pour les relatifs et les interrogatifs
-Cl pour les clitiques, y compris les clitiques sujets et l'adverbe de négation
-Pro pour les autres pronoms
-X pour les éléments dont on ne peut déterminer la catégorie syntaxique

- Quelles sont les catégories morpho-syntaxiques caractéristiques des genres discursifs, des types de conversations ? Générer un graphique de ventilation des unités caractéristiques (formes, segments, POS) sur la partition de votre choix.
- Contraster le discours interactif/non-interactif/semi-interactif par le calcul des spécificités POS (Part-Of-Speech). Générer une visualisation graphique type Partie−IndiceSpécif|FQ|fq
- Calculer les segments répétés du corpus au niveau de l'annotation morpho-syntaxique (POS). Quels sont les deux segments les plus fréquents ? Sont-ils caractéristiques d'un genre discursif particulier ? Les afficher dans une carte des sections et explorer les contextes.

**Exploration 2 (comprendre le marquage prosodique et la hiérarchie prosodique, fusionner les annotations) :**

- Créer une nouvelle couche d'annotation qui fusionne la catégorie (POS) et l'information concernant le début de la période intonative (IPE), qui est la macro-unité la plus grande au sein de la hiérarchie prosodique considérée dans le corpus.

Note : l'appartenance à une période intonative est donnée au format BILOU (**annotation n° 42**) :
B - 'beginning'
I - 'inside'
L - 'last'
O - 'outside'
U - 'unit'

- Identifier les POS et les segments répétés POS en début des périodes intonatives. Quelle est leur répartition par genres ?
- Combiner les résultats de l'analyse précédente avec le degré de proéminence initiale de la première syllabe du token (une proéminence peut être annotée W pour Weak ou S pour Strong, (*cf.* la description de l'**annotation n° 58**).
- Sur la base des analyses précédentes, étudier en contexte les éléments caractéristiques dont la saillance initiale et la proéminence varient en fonction de genres discursifs.
- Comment la saillance initiale reflète les besoins communicationnels (interaction, tours de parole, …) ?

**Exploration 3 (analyse des dépendances syntaxiques et le marquage prosodique : fonctions avancées, niveau « expert » :-) :**

- Quelles relations syntaxiques caractérisent les débuts des périodes intonatives ?

## Références bibliographiques

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.

Zimina M., Ballier N. (2018). "On the phraseology of spoken French: initial salience, prominence and lexicogrammatical recurrence in a prosodic-syntactic treebank Rhapsodie". Proceedings of JADT 2018: International Conference on Statistical Analysis of Textual Data, 12-15 June 2018, Rome, Italy.

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 of 13-14 November 2017, London: Computational and Corpus-based Phraseology: Recent Advances and Interdisciplinary Approaches. Volume II.