Extraction de Phrases Préfabriquées des Interactions à partir d'un corpus arboré du français parlé : une étude exploratoire

Marie-Sophie Pausé, Agnès Tutin, Olivier Kraif, Maximin Coavoux

Extraction de Phrases Préfabriquées des Interactions à partir d'un corpus arboré du français parlé : une étude exploratoire

Marie-Sophie Pausé[1,*], *Agnès* Tutin[1], *Olivier* Kraif[1], et *Maximin* Coavoux[2]
[1]Univ. Grenoble Alpes, LIDILEM, Bâtiment Stendhal CS40700 38058 Grenoble cedex 9, France
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, CS 40700 38058 Grenoble Cedex 9, France

Résumé. Dans cette étude exploratoire, nous nous intéressons aux Phrases Préfabriquées des Interactions (p. ex. *c'est clair ; je te jure ; on dirait*). Après avoir défini ce type de phrase, nous évaluons dans quelle mesure le corpus arboré Orféopeut être exploité pour extraire et caractériser ces éléments. Les résultats de l'analyse qualitative montrent que le repérage des phrases parenthétiques apparaît plus complexe que pour les clausatifs (propositions indépendantes). Nous montrons aussi comment l'outil Lexicoscope permet, en exploitant la combinatoire lexico-syntaxique et la distribution des éléments entre et à l'intérieur des tours de parole, de mieux cerner les caractéristiques de ces phrases préfabriquées.

Abstract. Extraction of Prefabricated Interaction Phrases from a French Spoken Treebank : An Exploratory Study. In this exploratory study, we are interested in Prefabricated Phrases of Interactions (e.g., 'c'est clair'; 'je te jure'; 'on dirait'). After defining this type of sentence, we evaluate to what extent the Orféo treebank can be exploited to extract and characterize these elements. The results of the qualitative analysis show that the identification of parenthetical sentences appears to be more complex and difficult than for clausatives. We also show how the Lexicoscope tool allows us, by exploiting the lexical-syntactic combinatorics and the distribution of the elements between and within the speech turns, to better identify the characteristics of these prefabricated sentences.

1 Introduction

Les travaux sur les interactions orales et écrites ont pris un réel essor ces dernières années. Néanmoins, le domaine de la phraséologie y reste peu représenté, malgré la présence importante d'expressions spécifiques dans ce type d'échanges et la forte demande en didactique des langues, où l'on souhaite de plus en plus exploiter des études et des matériaux basés sur des données authentiques. Le faible nombre de travaux descriptifs s'explique, d'une part, par le peu de corpus disponibles jusqu'à très récemment, et d'autre part, par la rareté des modèles intégrant tous les paramètres propres aux interactions, même si on peut mentionner des travaux récents prometteurs comme ceux de Blanco & Mejri (2018), Kauffer (2013 ; 2019) et Krzyżanowska *et al.* (2021). Les corpus nouvellement mis à disposition de la communauté scientifique permettent maintenant d'envisager des études plus approfondies sur les phénomènes phraséologiques propres aux interactions. Les problématiques au cœur des travaux actuels sont alors d'évaluer dans quelle mesure on peut exploiter ces nouveaux corpus à cette fin, et avec quels outils.

* Corresponding author : pauselinguist@gmail.com

L'objectif de la présente communication est d'évaluer l'intérêt d'utiliser des corpus arborés de l'oral pour l'extraction et l'étude des phrases préfabriquées des interactions. Pour ce faire, nous exploitons le corpus arboré Orféo[1] (Benzitoun & Debaisieux, 2020) qui est à ce jour le plus gros corpus arboré disponible pour le français parlé, présentant par ailleurs plusieurs genres de l'oral. En effet, ce corpus présente l'avantage d'avoir été analysé au plan syntaxique, ce qui permet d'envisager l'exploitation d'outils spécifiques pour le repérage et la caractérisation des phrases préfabriquées des interactions.

Après avoir défini dans une première section la notion de « phrase préfabriquée des interactions », nous détaillerons dans une seconde section le corpus adopté pour notre analyse. Nous présenterons ensuite une étude de cas de 7 phrases préfabriquées de l'oral, à partir de leurs caractéristiques interactionnelles et des analyses syntaxiques effectuées et évaluerons les difficultés rencontrées pour une future exploitation à plus grande échelle. Enfin, la dernière section présentera l'adaptation de l'outil Lexicoscope aux corpus oraux et proposera quelques pistes d'exploitation pour notre objet d'étude.

2 Les Phrases Préfabriquées des Interactions (PPI)

Notre objet d'étude se situe dans la phraséologie propre aux interactions. Nous définirons dans un premier temps la notion de « phrase préfabriquée des interactions », avant d'en préciser dans un second temps les caractéristiques principales.

2.1 Définition

L'émergence des travaux portant sur la spécificité des unités phraséologiques dans les interactions (Krzyżanowska *et al*., 2021 ; Blanco & Mejri, 2018)[2] a fait naître une terminologie propre à ce champ d'investigation : « phraséologie exclamative » (Bally, 1909), « routines conversationnelles » (Klein & Lamiroy, 2011), « énoncés liés » (Fónagy, 1997 ; Marque-Pucheu, 2007). Ce foisonnement terminologique peut être expliqué par la richesse des angles sous lesquels ce phénomène peut être observé. Nous adopterons ici en guise de terme-chapeau la dénomination « Phrases Préfabriquées des Interactions », désormais PPI, proposée par Tutin (2019). Il s'agit d'une notion permettant de regrouper différents types d'unités phraséologiques déjà décrits par ailleurs[3] dont le dénominateur commun est l'utilisation privilégiée au sein d'interactions orales ou écrites.

Les PPI font partie de la classe des unités phraséologiques et, à ce titre, présentent un fonctionnement qui ne peut pas être expliqué intégralement à partir des régularités syntaxiques et sémantiques de la langue (Burger *et al*., 1982 ; Schmale, 2013). Leurs caractéristiques sont les suivantes[4] :

a) elles sont sélectionnées en bloc par le locuteur ;
b) elles présentent des restrictions syntaxiques et lexicales ;
c) elles peuvent constituer à elles seules un acte illocutoire ;
d) elles apparaissent préférentiellement dans des interactions et leur interprétation est étroitement liée à la situation d'énonciation ;
e) des propriétés précédentes découle une difficulté à les traduire littéralement dans d'autres langues.

Prenons l'exemple (1).

(1) [L1] bon ben passe un bon dimanche toi [L2] eh ben merci toi aussi / et bon bon lundi surtout [L1] oh **tu parles** j'ai pas envie d'y aller hein [L2] oui ben ça ouais je sais / mais ça fait rien (C-Oral-Rom > ftelpv08)

La combinaison *tu parles* constitue à elle seule une unité illocutoire. Elle présente des contraintes lexicales et syntaxiques étant donné que, dans le même contexte, on ne pourrait pas avoir *vous parlez* ou *tu parlais* ou bien encore *tu causes*. Elle permet au locuteur d'exprimer une désapprobation en réaction aux propos de son interlocuteur (Krzyżanowska *et al*., 2021 : 516). Les traductions en polonais et italien proposées (*ibid*.) sont : *Jasne*! et *Ma scherzi?!*. Cette combinaison est une PPI, de la même façon que *dis donc, comment dirais-je, ça va pas la tête, que veux-tu que je te dise* ou *il n'y a pas de quoi*.

2.2 Caractéristiques formelles

Nous venons de présenter les principaux traits définitoires des PPI. Nous allons à présent donner des précisions sur leurs caractéristiques formelles et fonctionnelles. Nous commençons par la notion de « phrase » dont la définition est souvent problématique (Le Goffic, 2001). Nous l'utilisons ici non pas pour caractériser des phénomènes de syntaxe orale (Pietrandrea *et al*., 2014) mais bien pour décrire un phénomène phraséologique transversal aux canaux oral et écrit. L'objectif est d'aborder des unités phraséologiques transversales aux canaux écrit et oral. Ajoutons aussi que le terme « phrase » est plus facilement exploitable en didactique (où il y a une forte demande pour ce type de description) que des termes plus techniques. Les éléments que nous entendons par « phrases » sont des segments syntaxiques autonomes constitués d'un prédicat et assortis d'une modalité d'énonciation (Lefeuvre, 2007). On se rapproche ainsi de la notion initiale de phrase telle qu'on peut la trouver chez Furetière (1690). Le segment comporte ainsi une tête, dont la valence est complète, et constitue une unité illocutoire. La tête peut être verbale (par exemple, *tu parles*) ou *averbale* comme dans l'exemple 2.

(2) [L1] j'espère que c'était pas ce lundi-là parce qu'on n'a pas pu venir [L2] ah ah **la blague** (TUFS > 12_JG_AI_100224)

Parmi ces PPI, certaines constituent des propositions indépendantes, non reliées à d'autres phrases. Elles peuvent être intégrées à la classe des clausatifs de Mel'čuk (2006) et Milicevic (Mel'čuk et Milićević, 2014). *Tu parles* dans (1) ou *la blague* dans (2) seront ainsi considérés comme des clausatifs, car ces expressions constituent des phrases indépendantes, non subordonnées, constituant des actes illocutoires, ici principalement expressifs.

À côté des clausatifs, on observe une série de PPI qui présentent des propriétés différentes : elles sont facultatives au plan syntaxique, souvent déplaçables et sont insérées dans une autre phrase sans lien de subordination. Ces phrases « parenthétiques » sont des « unités qui pourraient être des unités illocutoires indépendantes mais elles se trouvent insérées dans une autre unité illocutoire » (Kahane & Gerdes, 2020 : 81 ; voir également Kahane & Pietrandrea, 2009). Sur le plan sémantique, ces phrases constituent souvent des commentaires méta-énonciatifs sur le dire ou le dit, ou fonctionnent comme des marqueurs d'interaction phatiques. Dans les exemples suivants, *comment dire* ou *tu sais* correspondent à des parenthétiques[5].

(3) [B] c'est une espèce de plante euh **comment dire** cérémoniale médicinale magique [A] médicinale (orféo oral, tufs)

(4) [STE] … il revient dans un dans son village natal pour al-~ enterrement d'un copain / et il déteste l'ambiance de son village natal [STE] c'est un monde ancien et tout **tu sais** (orféo oral, coralrom)

Cependant, la portée des deux parenthétiques n'est pas la même. Dans (3) *comment dire* porte sur un élément de la phrase, alors que dans (4) le parenthétique *tu sais* porte sur toute

la phrase dans laquelle il est inséré. En outre, certaines PPI connaissent une certaine variabilité comme des variantes temporelles ou l'insertion de modifieurs, comme nous le verrons à la section 4. Certaines PPI peuvent également avoir des arguments facultatifs comme dans *tu parles d'un cadeau.*

2.3 Caractéristiques fonctionnelles

À côté des propriétés syntaxiques, les PPI correspondent aux plans pragmatique et sémantique à plusieurs types d'unités phraséologiques déjà théorisés par les linguistes. On pourra ainsi évoquer les pragmatèmes de Mel'čuk (2013) et Blanco & Mejri (2018), ainsi que les formules expressives de la conversation de Krzyżanowska *et al.* (2021), les *speech formulae* de Cowie (2001), ou les actes de langage stéréotypés de Kauffer (entre autres, 2019). Sans entrer dans les détails, nous proposons de notre côté, à la suite de Tutin (2019), une typologie fonctionnelle provisoire des PPI, qui pourra être affinée (pour une typologie plus fine, *cf.* López Simó, 2016). Nous distinguons alors les types de PPI suivants :

- PPI méta-énonciative : porte sur le contexte d'énonciation, le dire et le dit ; *Tu vois, Tiens-toi bien* (*cf.* « marqueurs de discours propositionnels » (Andersen, 2007), « Formules métacommunicatives » López Simó (2016), « Métaphrases » (Dostie, 2019), « speech formulae » (Cowie, 2001)) ; *Si tu veux, Je crois ;Comment dirais-je, Je veux dire*
- PPI réactive : liée à une réaction (à la situation ou à l'interaction) ; *Il manquerait plus que ça, La blague, C'est malin* (*cf.* « Actes de langage stéréotypés » (Kauffer, 2013, 2029), « Formules expressives de la conversation » (Krzyżanowska *et al*., 2021)) ; *C'est clair, Comme tu veux, Je veux bien*
- PPI situationnelle (phrase épisodique, non réactive, dont l'interprétation référentielle est liée à la situation d'énonciation) : *Il y a de l'eau dans le gaz, Il n'y a pas grand-monde, Quand on parle du loup…* (*cf.* « Phrases situationnelles » (Anscombre, 2000))
- Pragmatèmes (phrases associées à des situations sociales ou communicatives contraintes et spécifiques) : *Ça va ?, Ça fait plaisir, Ça fait longtemps* (*cf.* Mel'čuk (2013), Blanco & Mejri (2018), « Phrases sociales » (Dostie, 2019))

Bien entendu, certaines PPI peuvent simultanément correspondre à plusieurs de ces types, comme le montre l'échantillon étudié à la section 4. En outre, comme nous le verrons plus loin, nombre de nos expressions comme *dis donc* sont polyfonctionnelles. Elles ont clairement plusieurs acceptions distinctes et peuvent se ranger sous plusieurs types. *Dis-donc* peut ainsi apparaître comme un réactif expressif communiquant l'étonnement ou l'incrédulité, ou comme une amorce pour aborder un nouveau thème dans la conversation.

3 Corpus et outils de traitement du corpus

Notre projet vise à extraire et à caractériser les PPI les plus productives des interactions du français. Dans cette perspective, nous avons choisi d'utiliser le corpus arboré Orféo (Benzitoun et Debaisieux, 2020), en repérant les PPI à l'aide de l'outil Lexicoscope, développé par O. Kraif au LIDILEM (2019)[6].

3.1 Le Corpus CEFC-ORFEO

Le corpus CEFC (*Corpus d'Étude pour le Français Contemporain*) a été développé dans le cadre du projet ANR ORFÉO (Outils et Recherche sur le Français Écrit et Oral, 2012-2016, Benzitoun & Debaisieux, 2020), et intègre un sous-ensemble représentatif de données textuelles écrites et orales, librement disponibles. La partie orale constitue, à notre connaissance, le corpus le plus volumineux et le plus diversifié de ce type, intégrant de nombreux échantillons d'autres corpus de référence[7] pour l'étude du français parlé :

conversations, entretiens, réunions diverses et discours publics, pour un total de 3 088 443 mots (*cf.* détails du corpus en fin d'article). Le corpus arboré Orféo – la partie orale du corpus – comporte des annotations syntaxiques en dépendances (Kahane & Gerdes, 2020 ; Deulofeu & Valli, 2020 ; Nasr *et al.,* 2020), dont une partie a été validée manuellement, et l'autre partie a été analysée automatiquement. Le schéma d'annotations comporte des annotations spécifiques à la syntaxe de l'oral, comme on le verra à la section 3. Au regard des phénomènes phraséologiques que nous voulons étudier, le corpus reste de petite taille. Néanmoins, il présente l'intérêt d'une diversité en termes de types d'interactions et l'annotation syntaxique est intéressante pour l'identification de structures lexico-syntaxiques récurrentes et de phénomènes représentatifs. Il nous est à ce titre nécessaire de vérifier l'exploitabilité du corpus global, et non seulement la partie annotée manuellement.

Le corpus est interrogeable en ligne soit par une interface simple, qui permet de sélectionner différentes métadonnées (types de corpus, nombre de locuteurs, etc.) mais ne permet pas d'exploiter l'annotation morphologique et syntaxique, soit par l'interface Grew-match (Guillaume, 2021) qui manipule de nombreux corpus arborés[8]. L'outil Grew-match offre des possibilités de requêtes multiples via un langage de requête ***expressif*** et élaboré, et il est également possible d'interroger les lemmes, les mots-formes ainsi que les liens de dépendances syntaxiques. Nous souhaitions de notre côté observer comment les syntagmes qui nous intéressent étaient annotés dans le corpus afin d'évaluer comment il était possible d'exploiter l'annotation syntaxique en place. L'idée était de partir des constituants lexicaux et d'observer les différentes structures lexico-syntaxiques qui leur étaient associées. Nous avons pour cela utilisé l'outil Lexicoscope, plus simple d'utilisation et mieux adapté à nos explorations que l'interface Grew Match (*cf.* en particulier section 4).

3.2 L'intégration d'Orféo dans le Lexicoscope

Le Lexicoscope[9] (Kraif, 2019) est une interface dédiée à l'exploration de corpus arborés et à l'étude de la combinatoire lexicosyntaxique. Tout comme le très populaire Sketch Engine[10], il s'appuie sur les dépendances syntaxiques pour extraire les cooccurrences et caractériser le voisinage des mots. Il permet notamment d'obtenir des statistiques variées ainsi que les concordances pour le retour au texte. Le concordancier a été adapté récemment pour intégrer l'affichage des locuteur·ices, des tours de parole et donner un accès aux enregistrements audios. L'outil a été utilisé dans de nombreux projets comme PhraseoRom ou TermITH pour l'extraction de la phraséologie à partir des lexicogrammes et de la méthode des Arbres Lexico-syntaxiques Récurrents (*cf.* section 5). Pour cette étude, nous avons intégré Orféo à l'interface du Lexicoscope.

A l'instar de Grew-match, l'outil permet d'effectuer des requêtes à partir d'un langage exploitant les formes, les lemmes, les traits morphosyntaxiques et les dépendances (langage TQL)[11]. Par exemple, pour la PPI *comme tu veux*, on utilisera la requête suivante :

```
<l=comme,c=CSU,#1>&&<l=tu,c=CLS,#2>&&<w=veux,c=VRB,#3>::(dep,1,3)(subj,3
,2)
```

Comme il est malaisé de composer soi-même ces requêtes, qui nécessitent de connaître à la fois le formalisme TQL mais aussi les jeux d'étiquettes et le découpage en tokens propre à chaque corpus, le système est doté d'un module de génération des requêtes : il suffit d'entrer une réalisation possible de l'expression pour voir s'afficher différentes suggestions proposées. Comme il n'est pas rare qu'une même expression reçoive différentes analyses (et corresponde donc à différentes requêtes), l'utilisateur·ice pourra ainsi couvrir toutes les occurrences du phénomène qu'il recherche. Par exemple, l'outil peut proposer à l'utilisateur

différentes structures lexico-syntaxiques associées à *comme tu veux* en cherchant dans le corpus des arbres de dépendances dont certaines positions sont occupées par les unités constitutives de la combinaison. L'utilisateur choisira ensuite quelle(s) requête(s) il veut lancer. Ce mode de requête est intéressant lorsqu'on souhaite explorer la qualité de l'annotation syntaxique.

Une fois la requête lancée, l'utilisateur peut afficher les concordances, sous format KWIC ou obtenir un contexte plus large (*cf.* figure 1). Il peut aussi écouter les segments où apparaît la requête (grâce à l'alignement texte-son effectué dans le corpus) ou visualiser les arbres syntaxiques associés.

Fig. 1: Extrait de la concordance KWIC de comme tu veux dans le CEFC-ORFEO

Un calcul de spécificité statistique (Kraif, 2019), permet également d'identifier si une expression est surreprésentée ou sous-représentée dans les différents sous-corpus. On observe ainsi que l'expression est plus courante dans l'oral conversationnel spontané, plus présent dans TCOF, CLAPI ou TUFS que dans des corpus d'entretiens un peu plus formels comme CFPP.

4 Étude de cas

L'objectif de la présente étude est de tester l'utilisabilité d'Orféo en intégralité pour l'extraction et la caractérisation des PPI. Il s'agit alors de vérifier si l'analyse syntaxique nous permet de repérer des emplois spécifiques aux PPI. On veut tout d'abord pouvoir discriminer les emplois comme PPI des constructions libres. Dans un second temps, on veut évaluer si l'analyse syntaxique permet de distinguer plusieurs emplois.

4.1 Échantillon analysé et repérage des différents emplois

Pour cette étude de cas, nous avons sélectionné un échantillon de PPI qui contiennent un verbe fléchi. Pour arrêter notre choix, nous avons retenu les critères de fréquence, de variabilité des structures et de présence au sein de plusieurs sous-corpus. Certaines PPI peuvent avoir plusieurs emplois appartenant à différents types syntaxiques. Ainsi, une même construction lexico-syntaxique peut être rencontrée comme clausatif ou comme parenthétique. L'échantillon, relativement restreint afin de privilégier une étude qualitative plutôt que quantitative, comporte 7 items rassemblés dans le Tableau 1[12].

Le tableau permet de voir qu'une même combinaison peut avoir deux emplois en tant que PPI différentes. C'est le cas de *je te jure* qui peut s'employer comme parenthétique ou comme clausatif, avec deux fonctions sémantico-pragmatiques différentes. Une même construction lexico-syntaxique peut également avoir deux fonctions sémantico-pragmatiques, comme les emplois parenthétiques de *c'est clair* et *je te jure*. Nous pouvons en effet les associer à la fois à une fonction méta-énonciative métalinguistique – ils portent sur le dit du locuteur – et à la fois à une fonction réactive expressive / évaluative – ils opèrent une évaluation du contenu référentiel du message.

Pour traiter l'ambiguïté de certaines PPI qui peuvent aussi être des constructions libres, nous avons procédé en deux étapes. La première étape a consisté à adapter la requête sur le Lexicoscope afin de réduire le bruit. Ainsi, la requête pour identifier les emplois de *je dirais* exclut en position de dépendant du verbe les adverbes de négation et les clitiques, afin de ne pas extraire *je le dirais* ou *je dirais pas*, par exemple. La seconde étape nous a amenés à opérer un tri manuel en fonction du contexte et de l'analyse syntaxique, pour éliminer d'autres occurrences qui n'étaient pas des PPI et pour discriminer les différents emplois des PPI[13]. Nous verrons en section 4.2 sur quels critères nous nous sommes appuyés.

Tableau 1: Échantillon de PPI étudié.

Phrase (nbr occurrences)	Exemples d'emplois comme PPI dans le corpus	Types syntaxique et sémantico-pragmatique de la PPI	Glose
comme tu veux (38 occ.)	[LAU] allez on y va / [JEA] **comme tu veux** hein / [JUL] attends-moi je vais juste éteindre les lumières (CLAPI > aperitif_chat)	clausatif ; réactive	Permet au locuteur, en réaction à un énoncé de son interlocuteur, de lui signifier qu'il s'adapte à sa volonté. Dans certains contextes, peut également permettre à son interlocuteur d'acquiescer en apparence pour mettre fin à une dissension.
je te jure (134 occ.)	[RL] ouais normal quoi / c'est Mac Gyver [AB] **je te jure** / ah purée je me souviens je regardais ça avec mon père parce que à l'époque on voulait faire Pékin Express (TUFS > 13_AB_RL_100224)	clausatif ; réactive	Permet au locuteur de réagir à son énonciation précédente ou à celle de son interlocuteur en exprimant un effarement
	[L2] il y en avait un qui s'était coincé le doigt d'ailleurs [...] **je te jure** la main énorme quoi mais le truc tu sais la main de de de Mickey (TCOF > laura)	parenthétique ; méta-énonciative & réactive	Permet au locuteur, en réaction à ses propres propos, de renforcer son énonciation.
on dirait (105 occ.)	[FV1] on dirait vercingétorix peut-être / [GU2] **on dirait** vous avez raison (CLAPI > visite_guidee_manoir_guide2)	clausatif ; réactive	Permet au locuteur, en réaction à un énoncé de son interlocuteur, de lui signifier qu'il observe la même chose que lui.
	[OG] sa tête j'aime pas sa gueule / [OG] je sais pas il est pointu du menton / [OG] **on dirait** / [OG] mais le reste est carré (tufs > 01_OG_NH_100222)	parenthétique ; méta-énonciative	Permet au locuteur de nuancer ses propos en indiquant qu'il s'agit de sa perception personnelle
dis donc (27 occ.)	[L2] c'est c'est un mot anglais [L1] oui c'est un mot anglais ouais F E E S [L2] **dis-donc** [L1] voilà [L2] je suis morte de rire (crfp PRO-PCR-1.orfeo)	clausatif ; réactive	Permet au locuteur, en réaction à un énoncé de son interlocuteur, de lui signifier son étonnement.
	dis donc tu sais la pièce qu'on avait l'autre jour d'allumage là pour l'A X euh on s'en est débarrassée (C-Oral-Rom > ffamdl01)	parenthétique (amorce) ; méta-énonciative	Permet au locuteur d'attirer l'attention de son interlocuteur
c'est clair (603 occ.)	[MB] à vie ils ont un emploi pour faire tous les côtés / [CT] **c'est clair** / (TUFS > 14_CT_MB_100224)	clausatif ; réactive	Permet au locuteur, en réaction à un énoncé de son interlocuteur, de renforcer l'énonciation de celui-ci.
	[SYL] c'est très bien hum / [SYL] **c'est clair** / (C-Oral-Rom > ffamdl03)	parenthétique ; méta-énonciative & réactive	Permet au locuteur, en réaction à ses propres propos, de renforcer son énonciation.
je [te/vous] dis (327 occ.)	ouais mais après tu sais **je te dis** les vieux italiens et les [sic] tu connais je te fais pas de dessins (reunions-de- travail > OF1_CA_3Dec07)	parenthétique ; méta-énonciative	Permet au locuteur de renforcer son énonciation en mettant l'accent sur une information qu'il a déjà énoncée.
je dirais (664 occ.)	[PAU] c'est c'est une grande idée en même temps qui va euh structurer **je dirais** euh euh et en tout cas la seconde moitié du dix-neuvième siècle voire même au-delà (CORALROM (O) > fnatco01)	parenthétique ; méta-énonciative	Permet au locuteur de nuancer un élément de son énoncé en indiquant qu'il s'agit de sa perception personnelle.

4.2 Observations sur l'analyse syntaxique

Après extraction des tris des occurrences de notre échantillon, nous avons établi une grille d'analyse afin d'observer l'analyse syntaxique intégrée au corpus.

4.2.1 Annotation Orféo

Le schéma d'annotation syntaxique adopté pour Orféo s'articule autour de la micro- et macro-syntaxe (Kahane & Gerdes, 2020). Les relations micro-syntaxiques vont nous intéresser préférentiellement pour la structure interne des PPI. Elles identifient les fonctions sujet (*suj*), spécifieur (déterminant, *spe*) , auxiliaire (*aux*) et dépendant autre (complément ou ajout, *dep*). Une relation pour les segments non analysables est également prévue (*disflink*), ainsi que des liens paradigmatiques (*para* et *mark*). L'analyse macro-syntaxique permet de décrire les éléments non régis au sein d'un énoncé. Sont identifiés :

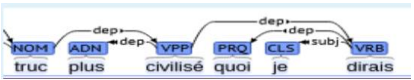
- composantes périphériques (*periph*)
- marqueurs de discours (*dm*) : éléments qui portent une forme de force illocutoire propre, prédiquant sur le noyau principal et n'acceptant pas de modifieurs
- parenthétiques (*parenth*) : unités ayant des propriétés illocutoires indépendantes mais se trouvant souvent insérées dans une autre unité illocutoire

4.2.2 Grille d'analyse

Notre grille d'analyse intègre trois paramètres principaux à étudier : la segmentation des énoncés, la syntaxe interne de la PPI et sa syntaxe externe, à savoir la façon dont elle est connectée au reste du segment dont elle fait partie. Il s'agissait alors de définir dans quelle mesure la segmentation était adaptée pour une analyse interne et externe correcte des PPI.

Pour illustrer la grille d'analyse adoptée, nous prendrons l'exemple d'une occurrence de *je dirais*. Le tableau 2 présente la description d'une occurrence de *je dirais*. Les segments intégrant la PPI sont surlignés en gris. À un segment correspond un arbre de dépendance syntaxique. Par manque de place, nous faisons apparaître seulement la partie de l'arbre qui contient la PPI et ses gouverneurs et/ou dépendants.

Tableau 2: Illustration de la grille d'analyse.

Source	crfp > PRI-NAR-1
Occurrence	L2] mais la prochaine fois j' irai plutôt dans un pays euh / [L2] comme la Guadeloupe ou la Martinique tu vois un truc plus civilisé quoi **je dirais** / [L1] hum hum hum ça dépend / [...]
Annotation syntaxique	
Position dans le segment	fin de segment
Syntaxe interne	*je* catégorisé comme clitique (CLS) et dirais catégorisé comme verbe (VRB), tous deux connectés par la relation syntaxique *sujet* (subj)
Syntaxe externe	La PPI est connectée au bon gouverneur syntaxique *civilisé.*

4.2.3 Syntaxe interne

La syntaxe interne des PPI observées — c'est-à-dire, la structure lexico-syntaxique de la PPI elle-même (voir Pausé, 2017, pp. 114-117) — est relativement régulière, puisque nous avons constaté seulement 2 cas de PPI sur les 7 où plusieurs annotations étaient possibles[14]. L'outil de suggestions de requêtes (*cf*. section 2.2) nous a en effet été utile pour *comme tu veux* et *dis donc*. Pour *comme tu veux*, la tête du syntagme est tantôt *vouloir*, tantôt *comme*. La seconde, *dis donc*, peut être identifiée sous deux structures syntaxiques : l'une avec un lien de dépendance *dep* entre *dis* et *donc*, et l'autre avec un lien *marqueur de discours (dm)*[15]. La relation *dep* relie un gouverneur à un dépendant qui occupe une fonction type complément d'objet ou circonstanciel.

L'analyse syntaxique permet également d'identifier l'ajout de satellites comme ***ah ben*** *c'est clair* ou **ça**, *c'est clair,* mais également des modifieurs comme *c'es**t bien** clair*. On peut également relever le cas d'insertion d'un élément tête facultatif comme dans ***(c'est)** comme tu veux*. Nous émettons l'hypothèse selon laquelle l'identification des satellites pourrait nous aider à identifier les différents emplois d'une même PPI (cf. 5 pour une méthode d'identification).

4.2.4 Segmentation et syntaxe externe

L'observation de la description syntaxique externe des PPI, à savoir la manière dont elles se combinent aux autres constituants d'un énoncé, nécessite de prendre en compte différents cas de figure. Nous distinguerons le cas des clausatifs de celui des parenthétiques.

Les PPI de notre échantillon qui ont un emploi clausatif sont, dans l'ensemble, bien analysées à partir du moment où il n'y a pas d'erreur de segmentation. Notons toutefois que les PPI comme *je te jure*, *c'est clair* et *on dirait*, qui peuvent à la fois être analysées comme clausatifs et parenthétiques présentent une analyse hétérogène. Pour les PPI de ce type, la position du segment qui contient la PPI par rapport aux changements de tours de parole est centrale pour son interprétation. Précisons qu'un tour de parole peut contenir plusieurs segments et que ces derniers sont séparés par le signe /.

(5) [AB] j' ai vu ça / [AB] **je te jure** / (tufs > 13_AB_RL_100224)
(6) [CD] ah oui en une demi-heure **je te jure**/ (tufs > 03_MW_CD_100222)
(7) [AM] vingt euros / [GL] c' est pas vrai / [AM] **je te jure**/ (tufs > 02AMGL110912)

Dans l'énoncé (7), la segmentation et le changement de locuteur d'un segment à l'autre permettent d'analyser *je te jure* comme un clausatif et une PPI réactive. Les énoncés (5) et (6) contiennent tous deux un seul tour de parole puisqu'un seul locuteur, mais la PPI *je te jure*, qui semble relever du même emploi, apparaît dans un nouveau segment dans (5), contrairement à 10. On voit là deux analyses différentes d'une même PPI. Ceci reflète la double catégorisation des PPI de ce type (*cf*. tableau 1), qui peut rendre leur repérage difficile en corpus.

Quand elles se trouvent dans un segment intégrant une autre proposition, les verbes des PPI comme *je te jure*, *c'est clair* et *on dirait* peuvent gouverner un autre constituant par la relation *dep* ou bien être dépendants d'un autre constituant via la relation *parenth*. Ils peuvent aussi être reliés à un autre verbe par la relation *para* qui indique un lien paradigmatique (type coordination), comme dans la figure 2.

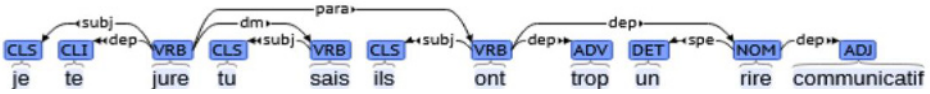


Fig. 2 : annotation de *je te jure*

À titre d'exemple, l'analyse de *je te jure* en position initiale d'un segment fait apparaître pour 41 occurrences sur 46 *jure* en tête, gouvernant une autre proposition par la relation *dep* dans 21 cas, par la relation *para* dans 18 cas et par la relation *dm* dans 2 cas. Dans 5 occurrences, *jure* est dépendant d'un autre élément par une relation *dep*, *para*, *parenth* ou *périph*.

Lors de l'observation détaillée des PPI uniquement parenthétiques, nous avons noté un nombre important de différences de segmentation probablement liées à l'analyse automatique (par exemple, 63 erreurs sur 175 occurrences de *je te dis*) qui entraîne inévitablement des erreurs d' analyse syntaxique des combinaisons qui nous intéressent.

(8) [L2] euh donc euh en roman alors j' ai euh l' avantage / [L2] donc euh **je vous dis** du gallo-roman c' est que on a les plus vieux atlas qui aient jamais existé (crfp > PRI-PRI-2)

Dans l'exemple 8, *je vous dis* est un parenthétique qui porte sur *l'avantage du gallo roman c'est qu'on a les plus vieux atlas qui aient jamais existé*. La segmentation opérée sur le tour de parole sépare *l'avantage* de « du gallo-roman » dont il est la tête et rend l'analyse syntaxique erronée.

Lorsque les énoncés sont bien segmentés, on note deux grands cas de figure :

a) la tête de la PPI est racine du segment
b) la tête de la PPI est un dépendant

Dans le cas a), le plus courant, le verbe de la PPI est considéré comme gouverneur syntaxique (ce qui n'est pas l'analyse attendue), lié à un dépendant par la relation *dep*. Dans la figure 3, par exemple, *je te dis* est annoté comme tête de la phrase.

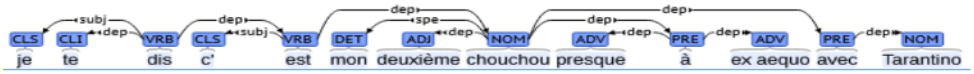


Fig. 3: annotation de la PPI *je te dis* en tant que gouverneur syntaxique de l'élément sur lequel elle porte.

Les annotations de ce type peuvent être exploitées dans la mesure où elles présentent une régularité. Un autre indice exploitable pour repérer les parenthétiques pourra être le paramètre de la position de la PPI au sein du segment. Dans l'exemple 9, la position de *je dirais* en fin de segment nous aiguille vers un emploi parenthétique.

(9) [ileFN1] ça je je je je ne signe jamais ce genre de chose parce que je trouve ça déplorable / c' est ça / mm / mm / d' utiliser ça / parce que c' est trop facile / c' est téléphoné un peu **je dirais**
(Valibel > ileFN1r)

Dans le cas b), la PPI est reliée à un gouverneur syntaxique — qui peut être celui attendu, ou non — par une relation *dm* ou *periph*. Dans la figure 4, *je dirais* est annoté comme marqueur de discours de *c'est obligé*.

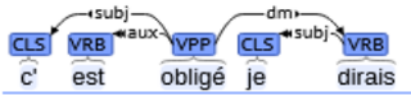


Fig. 4: annotation de *je dirais* comme marqueur de discours.

Comme son nom l'indique, la relation *periph* relie, dans l'annotation Orféo, la tête de la phrase à des éléments périphériques, à savoir tous les constituants situés à gauche du sujet, ainsi que les dépendants d'un verbe qui sont à la périphérie de la construction régie par celui-ci (toutes positions confondues). Les marqueurs de discours (*dm*) sont des types éléments périphériques qui ont la particularité d'être plus mobiles. Un *dm* est rattaché à l'élément qui le précède directement, ou à la racine de la phrase s'il est en position initiale du segment analysé.

Le cas de figure b) est le plus adapté pour l'extraction des PPI, car les liens de dépendance *dm* et *periph* sont en adéquation avec le fonctionnement syntaxique des PPI parenthétiques. Néanmoins, la faible fréquence des annotations de ce type nous interroge sur son exploitabilité dans le cadre de l'extraction de PPI à grande échelle. À titre d'exemple, sur 113 occurrences de *je te dis* en tant que PPI parenthétique, 100 annotations relèvent du cas a) introduit ci-dessus (PPI reliée à l'élément sur lequel elle porte par la relation *dep*), contre 13 annotations correspondant au cas b) (relation *dm* ou *periph*).

En synthèse, pour le repérage des clausatifs, l'annotation syntaxique est facilement exploitable lorsqu'ils constituent un segment à part entière et qu'il y a changement de locuteur par rapport au segment précédent. Si le locuteur reste identique d'un segment à l'autre ou bien si le segment contient d'autres propositions, l'occurrence doit être analysée plus en détail[16]. Pour les parenthétiques, les relations *dm* et *periph* peuvent nous aider à repérer les PPI. Malheureusement elles sont peu représentées, comme nous l'avons dit plus haut. On peut alors utiliser, comme nous l'avons vu un peu plus haut, les annotations qui intègrent le candidat PPI comme gouverneur syntaxique d'un élément via la relation *dep,* si la construction échappe à la combinatoire habituelle de la tête de la PPI.

5 Perspectives pour l'analyse et l'extraction des PPI à l'aide du Lexicoscope

Le profil combinatoire d'une expression pivot est rendu observable par l'extraction d'un « lexicogramme », un tableau indiquant les cooccurrents les plus fortement associés à l'expression. Par exemple, pour la requête suivante :

```
<l=ce,c=CLS,#1>&&<w=,c=VRB,#2>&&<l=clair,c=ADJ,#3>::(dep,2,3) (subj,2,1)
```

correspondant à la PPI *c'est clair*, on obtient la représentation illustrée par la figure 5.

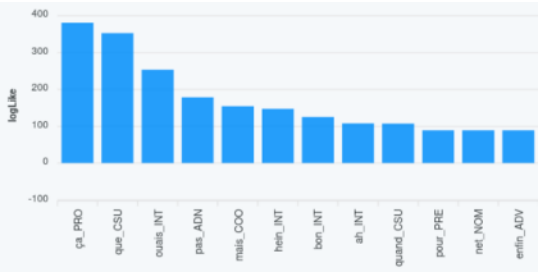


Fig. 5 : Extrait d'un lexicogramme pour la PPI *c'est clair.*

De la sorte, on identifie différents types de cooccurrents :

a) des cooccurrents qui indiquent des réalisations qui ne correspondent pas à la PPI, comme la conjonction *que* ou l'adverbe de négation, qui indiquent d'autres interprétations.

Connaître ces cooccurrents permet notamment de désambiguïser les requêtes pour mieux cibler l'expression cherchée. Par exemple, on peut affiner la requête en ajoutant des conditions négatives (en gras) :

```
<l=ce,c=CLS,#1>&&<l=être,c=VRB,#2>&&<l=clair,c=ADJ,#3>::(dep,2,3)
(subj,2,1)(!dep,2,<l=pas>)(!dep,3,<l=que>)
```

afin de désambiguïser et d'éliminer les occurrences de *c'est clair que...* (relation *dep* entre *clair* et *que*) ou « c'est pas clair » (relation *dep* entre *est* et *pas*). De même, on pourra spécifier que l'adjectif *clair* n'accepte pas d'adverbe dans ses réalisations dans la PPI (comme dans *c'est beaucoup plus clair*)

b) des cooccurrents qui indiquent des usages spécifiques de la PPI. Par exemple, les réalisations *ça c'est clair* ou *ouais c'est clair* indiquent en général une PPI de type clausatif réactif. Ces cooccurrents constituent une extension possible de la PPI.

L'application réitérée et automatisée de l'extraction des cooccurrents syntaxiques permet d'extraire ce qu'on appelle des Arbres Lexico-syntaxiques Récurrents (ou ALR, *cf.* Kraif *et al.*, 2014). Ces ALR permettent d'identifier en bloc des expressions polylexicales ou des constructions dans lesquelles l'expression cible vient s'insérer. Par exemple, autour de *c'est clair*, on peut extraire les ALR <INT mais c'est clair> (24 occurrences) ou <ah INT c'est clair> (18 occurrences) ou <ouais ça c'est clair> (6 occurrences) (INT désignant la catégorie des interjections).

Ces ALR peuvent être regroupés autour de constructions plus larges comme <INT INT c'est clair>. Partant d'une telle requête, le Lexicoscope permet d'afficher ses différentes réalisations, et de montrer les combinaisons d'interjections les plus fréquentes autour d'une telle PPI clausative réactive expressive. On observe ainsi les réalisations suivantes, par fréquence décroissante : "ah_INT oui_INT ce_CLS être_VRB clair_ADJ" (13), "ah_INT ouais_INT ce_CLS être_VRB clair_ADJ" (12), "ah_INT non_INT ce_CLS être_VRB clair_ADJ" (8), "quoi_INT ouais_INT ce_CLS être_VRB clair_ADJ" (5).

Si l'on cherche des expressions équivalentes ou proches de l'expression pivot, le Lexicoscope permet de rechercher des sous-arbres de même structure constitués à partir d'unités lexicales sémantiquement voisines (la similarité étant calculée à partir des *plongements lexicaux* de FastText[17], cf. Grave *et al.*, 2018). Cette « généralisation » permet

d'obtenir, dans le cas présent, les réalisations suivantes, qui partagent des propriétés avec la PPI *c'est clair*, tout en apportant des nuances sémantiques différentes : "ah_INT oui_INT ce_CLS être_VRB vrai_ADJ" (100), "ah_INT oui_INT ce_CLS être_VRB sûr_ADJ" (29), "ah_INT ouais_INT ce_CLS être_VRB bizarre_ADJ" (7), "ah_INT oui_INT ce_CLS être_VRB intéressant_ADJ" (7), etc. Ces expressions indiquent à la fois des quasi-synonymes (*c'est sûr*), des usages similaires (*c'est vrai*), mais aussi d'autres types réactifs expressifs (*c'est intéressant*, *c'est bizarre*) manifestant d'autres affects (étonnements, intérêt, ...).

Enfin, notons que le Lexicoscope permet également de préciser le positionnement des expressions cherchées à différentes échelles : début/milieu/fin de phrase, de tour de parole ou de document. Ainsi avec la requête suivante, l'ajout de la contrainte `paraPos=1` permet de filtrer les occurrences de l'expression en début de tour de parole, qui correspondent majoritairement à un emploi réactif :

```
<l=ce,paraPos=1,c=CLS,#3>&&<l=être,c=VRB,#1>&&<l=clair,c=ADJ,#5>::(dep,1,5)
(subj,1,3)
```

Si l'on veut privilégier les usages parenthétiques méta-énonciatifs, on pourra chercher des expressions qui apparaissent en fin de segment (sentPos=5) et ceci pour des segments comportant plus de 10 tokens ($sentLength>10) :

```
<l=ce,c=CLS,#3>&&<l=être,c=VRB,#1>&&<l=clair,sentPos=5,c=ADJ,#5>::(dep,1,5)
(subj,1,3)::$sentLength>10
```

6 Conclusion

Notre étude à visée qualitative nous a permis de vérifier l'exploitabilité de l'annotation syntaxique du corpus arboré Orféo pour l'extraction et la caractérisation des phrases préfabriquées des interactions. Nous avons pu mettre en exergue des paramètres importants comme :

- la syntaxe interne pour l'identification des modificateurs d'une PPI
- l'impact de la segmentation, la position de la PPI au sein du segment ainsi que les changements de tours de parole
- la nature de la relation connectant certaines PPI parenthétiques à leur gouverneur : marqueur discursif ou élément périphérique
- l'exploitation possible des cooccurrents (éléments satellites) pour la désambiguïsation.

Ces paramètres nous ouvrent des perspectives pour identifier de nouvelles PPI de manière semi-automatique à l'aide du Lexicoscope. L'objectif d'une extraction à grande échelle est aussi d'affiner la classification formelle et fonctionnelle des PPI.

Le corpus que nous avons utilisé, bien qu'il s'agisse actuellement du plus grand corpus arboré de l'oral disponible pour le français, reste relativement limité pour estimer la fréquence de PPI, notamment de celles qui sont peu fréquentes (pas ou peu d'occurrences dans Orféo). Par ailleurs, l'analyse en section 3.3. a mis au jour quelques incohérences liées aux erreurs de l'analyse automatique en dépendances réalisée sur la partie du corpus qui n'a pas été vérifiée manuellement. Pour pallier ces deux limites, un des prolongements de cette étude (en cours de réalisation) consiste en la préparation (analyse syntaxique automatique, lemmatisation) d'autres corpus. En particulier, nous utilisons des outils de l'état de l'art actuel pour l'analyse syntaxique en dépendances du français (Grobol & Crabbé, 2021) pour analyser l'ensemble du corpus ESLO 2 (Abouda & Baude, 2006) ainsi que la partie française du

corpus de sous-titres Opensubtitles (Lison & Tiedmann, 2016)[18]. Nous avons par ailleurs produit une autre version annotée du CEFC avec ces mêmes outils, en utilisant les jeux d'étiquettes UD. Malgré les défauts de ce dernier corpus pour notre étude (manque de métadonnées, données issues principalement de fictions), il a l'avantage d'avoir une taille suffisante (400 millions de tokens) pour appliquer nos méthodes d'extractions de PPI candidates ou pour estimer les fréquences de PPI en corpus.

Références bibliographiques

Abouda, L., Baude, O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO. In Rastier, F., Ballabriga, M. (dir.), Corpus en lettres et sciences sociales : des documents numériques à l'interprétation, Paris : *Texto*, pp. 143-50.

Bally, Ch. (1909). *Traité de stylistique française*. Paris : Klincksieck.

Benzitoun, C., Debaisieux, J. M. (2020). Orféo: un corpus et une plateforme pour l'étude du français contemporain. *Langages*, 219(3).

Blanche-Benveniste, C. (1989). Constructions verbales « en incise » et rection faible des verbes. *Recherches sur le français parlé,* 9, pp. 53-73.

Blanco, X. (2015). Les pragmatèmes: définition, typologie et traitement lexicographique. *Verbum*, 4(4), pp. 17-25.

Blanco, X., Mejri, S. (2018). *Les pragmatèmes*. Paris : Champion.

Bonami, O., Godard, D. (2008) Syntaxe des incises de citation. *Actes du premier Congrès Mondial de Linguistique Française*.

Britt, E., Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk*, 20(1), pp. 29-62.

Burger, H., Buhofer, A., Sialm, A. (1982). *Handbuch der Phraseologie*. Berlin/New-York : De Gruyter.

Cowie, A.P. (2001). Speech formulae in English : problems of analysis and dictionary treatment. *Groninger Arbeiten zur germanistishen Linguistik*, 44.

Deulofeu, J., & Valli, A. (2020). Lexique et classement en parties du discours dans ORFÉO, *Langages*, 219(3), pp. 53-68.

Dostie, G. (2019). Paramètres pour définir et classer les phrases préfabriquées : La vengeance est un plat qui se mange froid. Bon appétit! *Cahiers de lexicologie*, 114, pp. 27-61.

Fónagy, I. (1997). Figement et changement sémantique. Dans M. Martins-Baltar (dir.), *La locution entre langue et usages*. Fontenay-Saint-Cloud : ENS Éditions, pp. 131-164.

Furetière, A. (1690). *Dictionnaire universel. La Haye*. URL : https://gallica.bnf.fr/ark:/12148/bpt6k50614b/f1471.vertical

Granger, S., Paquot, M. (2008). Disentangling the Phraseological Web. Dans Sylviane Granger et Fanny Meunier (dir.), *Phraseology : an Interdisciplinary Perspective.* Amsterdam/Philadelphia : John Benjamins, pp. 27-49..

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T. (2018). Learning Word Vectors for 157 Languages, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Grobol, L., Crabbé, B. (2021). Analyse en dépendances du français avec des plongements contextualisés. *28e Conférence sur le Traitement Automatique des Langues Naturelles*, Jun 2021, Lille (virtuel), France.

Guillaume, B. (2021). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics*, Apr 2021, Kiev/Online, Ukraine.

Kahane, S., Gerdes, K. (2020). Annotation syntaxique du français parlé: Les choix d'ORFÉO, *Langages*, 219(3), pp. 69-86.

Kahane, S., Pietrandrea, P. (2009). Les parenthétiques comme Unités illocutoires associées. Une perspective macrosyntaxique, *Linx*, 61, Presses Universitaires de Paris Nanterre, pp. 49-70.

Kauffer, M. (2013). Le figement des « actes de langage stéréotypés » en français et en allemand », *Pratiques : théories, pratique, pédagogie*, 159-160, pp. 42-54.

Kauffer, M. (2019). Les "actes de langage stéréotypés" : essai de synthèse critique. *Cahiers de lexicologie*, 114, pp. 149-171.

Kavka, S., Zybert, J. (2004) Glimpses on the history of idiomaticity issues. *SKAZE Journal of Theoretical Linguistics*, 1, pp. 54-66.

Klein, J.-R., Lamiroy, B. (2011). Routines conversationnelles et figement. Dans J.-C. Anscombre et S. Mejri (dir.), *Le figement linguistique : la parole entravée*. Paris : Honoré Champion, pp. 195-214.

Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le LEXICOSCOPE. *Langue française*, 3, pp. 67-82.

Kraif, O., Tutin A., Diwersy S. (2014). Extraction de pivots complexes pour l'exploration de la combinatoire du lexique : une étude dans le champ des noms d'affect, *Actes du Congrès Mondial de Linguistique Française 2014*, 19-23 juillet 2014, Berlin, pp. 2663-2674.

Krzyżanowska, A., Grossmann, F., Kwapisz-Osadnik, K. (2021). *Les formules expressives de la conversation Analyse contrastive: français-polonais-italien*. Lublin : Episteme.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J., Obin, N., Pietrandrea, P., Tchobanov, A. (2014). « Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French ». *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA), p.295–301.

Le Goffic, P (2001). La phrase « revisitée », *Le français aujourd'hui*, 135(4), pp. 96-107.

Le Pesant, D. (2013). Syntaxe des introducteurs de Discours Rapporté au style direct. Dans S.Grosse, A.Hennemann, K.Plötner et S.Wagner (dir.), *Angewandte Linguistik*. Bern, Berlin : Peter Lang.

Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož : European Language Resources Association (ELRA), p.923-929.

López Simó, M. (2016*). Fórmulas de la conversación. Propuesta de definición y clasificación con vistas a su traducción español-francés, francés-español*. Thèse de doctorat. Université d'Alicante.

Mel'čuk, I. (2013).Tout ce que nous voulions savoir sur les phrasèmes, mais ..., *Cahiers de Lexicologie*, 102, pp. 129-149.

Marque-Pucheu, C. (2007). Les énoncés liés à une situation: mode de fonctionnement et mode d'accès en langue 2, *Hieronymus,* I, pp. 25-48.

Mel'čuk, I. (2006). Parties du discours et locutions, *Bulletin de la Société de linguistique de Paris*, 101, pp. 29-65.

Mel'čuk, I., Milićević J. (2014). *Introduction à la linguistique*, 2, Paris : Hermann.

Nasr, A., Dary, F., Bechet, F., & Fabre, B. (2020). Annotation syntaxique automatique de la partie orale du ORFÉO, *Langages*, 219(3), pp. 87-102.

Pausé, M. S. (2017). *Structure lexico-syntaxique des locutions du français et incidence sur leur combinatoire*. Thèse de Doctorat, Nancy, Université de Lorraine.

Pietrandrea, S., Kahane, A., Lacheret, F. Sabio (2014). The notion of sentence and other discourse units in spoken corpus annotation. Dans Mello H. , Raso, T. (dir.) *Spoken corpora and Linguistic Studies.* Amsterdam/Philadelphia : John Benjamins Publishing Company, pp. 331-364.

Rubio, E. (2020). Spanish phraseology in formal and informal spontaneous oral language production. *Yearbook of Phraseology*, 11. Berlin/New-York : De Gruyter, pp. 81-106.

Schmale, G. (2013). Qu'est-ce qui est préfabriqué dans la langue ? – Réflexions au sujet d'une définition élargie de la préformation langagière, *Langages*, 189, pp. 27-45.

Tutin, A. (2019). Phrases préfabriquées des interactions: quelques observations sur le corpus CLAPI. *Cahiers de Lexicologie*, 114, pp. 63-91.

Corpus du CEFC-ORFEO :

CFPB : https://www.corpusfinder.ugent.be/cfpb [consulté le 06/01/22]

CFPP : http://cfpp2000.univ-paris3.fr/ [consulté le 06/01/22]

CLAPI : http://clapi.ish-lyon.cnrs.fr/ [consulté le 06/01/22]

CORALROM : Cresti, E., & Moneglia, M. (dir.). (2005). *C-ORAL-ROM: integrated reference corpora for spoken romance languages* (Vol. 15). John Benjamins Publishing.

CRFP : Equipe Delic (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé*, *18*, 11-42.
French Oral Narrative : http://frenchoralnarrative.qub.ac.uk/ [consulté le 06/01/22]
Fleuron : https://fleuron.atilf.fr/ [consulté le 06/01/22]
OFROM : http://www11.unine.ch/ [consulté le 06/01/22]
TCOF : https://www.cnrtl.fr/corpus/tcof/ [consulté le 06/01/22]
TUFS :http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.htmlcontents_xml=top&menulang=en
[consulté le 06/01/22]
VALIBEL : https://uclouvain.be/fr/instituts-recherche/ilc/valibel/corpora.html [consulté le 06/01/22]

[1]https://www.ortolang.fr/market/corpora/cefc-orfeo
[2]Pour des travaux sur la fréquence d'emploi des unités phraséologiques en fonction des canaux écrits et oraux on pourra consulter Gutiérrez Rubio (2020) et Erman & Warren (2000).
[3]Pour un regard historique sur la phraséologie, voir notamment Kavka & Zybert (2004) et Granger & Paquot (2008)
[4]Notons que les PPI s'insèrent dans une classe plus globale de Phrases Préfabriquées dont Dostie (2019) donne les caractéristiques définitoires.
[5]Précisions que nous réutilisons le terme *parenthétique* pour caractériser la forme et le comportement des PPI qui ne sont pas clausatives. Notre usage ne correspond alors pas à la fonction syntaxique utilisée dans le cadre de l'annotation syntaxique d'Orféo (Kahane & Gerdes 2020).
[6]Les corpus de l'écrit (dialogue romanesques ou interactions écrites de discussions) seront aussi exploités.
[7]Les références vers ces corpus sont données en fin de bibliographie.
[8]http://match.grew.fr ; l'outil est également téléchargeable.
[9]http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0
[10]Le *Sketch Engine* est un outil développé par la société Lexical Computing, fondée par A. Kilgarriff, très utilisé dans le monde de la linguistique de corpus. Accessible à l'adresse : https://app.sketchengine.eu
[11]Ce formalisme est assez proche de CQL, le langage utilisé dans *Sketch Engine* et dans plusieurs autres outils, mais avec l'ajout de contraintes de dépendances. La documentation est accessible à : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/doc/Reference%20TQL.pdf .
[12]Notons que nous avons pris en compte l'alternance tutoiement/vouvoiement qui induit des variations flexionnelles. Nous avons par ailleurs écarté les formes du type *je te le dis* pour ne garder que la stricte combinaison de *dire* au présent de l'indicatif avec un sujet *je* et un clitique *te/vous*. Nous excluons également les occurrences de *je te dis* comme introducteurs de discours direct ou indirect. Ces derniers constituent un champ d'étude à eux-seuls (Le Pesant 2013, Bonami & Godard 2008).
[13]Par exemple, dans « avant on disait à l'anglaise maintenant on dirait plutôt américain », *on dirait* n'est pas une PPI, mais une construction libre du verbe *dire*.
[14] Nous pouvons également citer un problème de lemmatisation repéré pour *je te jure* que l'on peut trouver avec *jure* étiqueté comme un verbe à l'infinitif dans deux occurrences.
[15]La combinaison *dis donc* est également analysée, dans le corpus, comme mot-forme : l'adverbe *dis-donc*.
[16]La prosodie joue également un grand rôle, cependant le seul corpus dont nous ayons connaissance qui soit annoté en syntaxe et en prosodie (Rhapsodie, CITE) n'est pas de taille suffisante pour ce type d'études.
[17]Les plongements lexicaux sont des représentations vectorielles qui permettent d'assimiler les mots à des points dans un espace multidimensionnel. Les représentations sont calculées pour que des mots ayant des distributions similaires apparaissent dans des régions voisines de cet espace. La similarité sémantique peut alors être calculée géométriquement comme un cosinus entre vecteurs. Les plongements doivent être calculés sur de grandes quantités de textes, c'est pourquoi le Lexicoscope utilise des vecteurs génériques de FastText, qui sont disponibles pour de nombreuses langues.
[18]https://opus.nlpl.eu/OpenSubtitles-v2018.php