Studii de lingvistică 13, nr.2, 2023, 105 - 122

Aperçu d'une typologie de phrases préfabriquées des interactions en français dans des corpus parlés

Some Insights on a Typology of French Interactional Prefabricated Formulas in Spoken Corpora

Marie-Sophie Pausé[1]

Abstract: This study deals with interactional prefabricated formulas in French, the usual formulae used in spoken or written interactions, such as *comment dire?* 'how shall I put it?'; *tu parles!* 'you bet!'; *à tout à l'heure* 'see you later'; *de rien* 'you're welcome'; *content de te voir* 'nice to see you'; *c'est bien* 'that's fine'. These expressions, often called "routine formulae", have not been much listed and studied, unlike other types of expressions, although they are very common. We propose a broad typology of these formulas from meta-enunciative formulas (*on va dire* 'let's say') to direct interactional expressions (*c'est bon* 'it's OK') or ritual formulas (*bonne journée* 'have a nice day'). An annotation scheme has been developed to account for the different aspects of these elements including main types, semantic labels, clause types and lemmatization.

Keywords: phraseology, interactions, annotation, corpora

1. Introduction

Si les travaux sur les interactions se sont considérablement développés ces dernières années, notamment dans les domaines de la syntaxe et de la pragmatique interactionnelle, peu d'études ont été consacrées au lexique ou à la phraséologie interactionnelle. Il existe pourtant un ensemble d'expressions récurrentes, propres aux interactions orales et écrites, qui méritent d'être identifiées et étudiées dans le cadre de la phraséologie, telles que *on parie ?* ; *à plus tard* ; *de rien* ; *ravi de vous rencontrer* ; *c'est bon*. Ces expressions ont souvent été décrites comme des *routines conversationnelles* (Coulmas 1979) ou des *formules de routine* (Cowie 2001). Récemment, plusieurs études ont été consacrées à ce sujet (Blanco & Mejri 2018, Kauffer 2013 ; 2019 ; Krzyżanowska et al. 2021). Cependant, un large inventaire et

[1] pauselinguist@gmail.com.

une classification de ces expressions restent à faire, en particulier pour le français. Pour mieux explorer ce domaine, nous avons entrepris un travail empirique d'annotation sur des corpus oraux, afin d'identifier les unités phraséologiques interactionnelles et de les classer[2].

Nous proposerons d'abord une définition des Phrases Préfabriquées des Interactions avec un bref état de l'art. Nous présenterons ensuite le schéma d'annotation, incluant plusieurs paramètres.

2. Phrases Préfabriquées des Interactions

2.1. Définition

Cette étude traite des Phrases Préfabriquées des Interactions (ci-après PPI) ; ces formules habituellement utilisées dans les dialogues parlés ou écrits, plus que dans les discours monologués, telles que : *on parie ?* ; *à plus tard* ; *de rien* ; *ravi de vous rencontrer* ; *c'est bon*. Les dialogues courants sont pleins de ces formules préfabriquées, comme on peut l'observer dans l'exemple (1) (les PPI sont soulignées).

(1) [spk3] bah vous avez dû voir ça dans les /
[spk4] tu sais quand tu es y a le côté officiel / euh / ouais / euh moi je me suis mise à bégayer d' ailleurs je suis / un BTS euh p- /
[spk3] moi j'ai laissé parler Jennifer /
[spk4] ouais moi aussi oh vas - y /
[spk3] non mais ouais /
[spk4] hm /
[spk4] non mais c'est cool
[spk4] bon bah on va peut-être vous laisser ? / euh / ouais
[spk1] bah merci euh / bah bonne soirée / bon courage euh pour votre licence / merci / merci / merci beaucoup / au revoir [ESLO2 CINE_1187]

Ces éléments phraséologiques ont la particularité, comme les proverbes, d'être des clauses complètes, avec une valeur illocutoire, qui peuvent être utilisées de façon autonome. Ces formules ont été explorées sous différentes terminologies, de la *phraséologie exclamative* de Bally (1921 [1951]), aux *routines conversationnelles* de Coulmas (1979, 1985), Corpas Pastor (1996) ou Lüger (2007), en passant par les *énoncés liés* de Fónagy (1982) et Martins-Baltar (1995), avec une extension légèrement variable d'un auteur à l'autre.

[2] Les travaux présentés dans cet article ont été menés avec Agnès Tutin et en partenariat avec Olivier Kraif et Maximin Coavoux. Une version de ces travaux a été publiée en anglais dans le cadre du colloque Europhras 2022 : Pausé, Marie-Sophie et Tutin, Agnès, "Some insights on a typology of French interactional prefabricated formulas in spoken corpora", in Corpas Pastor, G. et Mitkov, R. (eds), *Computational and Corpus-Based Phraseology*, 4th International Conference Europhras 2022, Malaga, Spain.

Dans cette étude, nous attribuons aux PPI les propriétés suivantes :

a) Ce sont des clauses complètes, avec une valeur illocutoire, même si elles peuvent être enchâssées en tant que propositions subordonnées ou parenthétiques, comme *comment dirais-je* dans l'exemple (2) ci-après.

b) Elles sont récurrentes dans les interactions. La plupart d'entre elles sont fréquentes dans les dialogues.

c) Ce sont des formules toutes faites. L'utilisation de beaucoup d'entre elle est fortement déterminée par le contexte. Pour répondre poliment à un merci, un francophone peut choisir une formule parmi la liste suivante : *de rien, il n'y a pas de quoi, ou tout le plaisir est pour moi.*

d) La plupart d'entre elles ne peuvent être interprétés littéralement et ne sont pas sémantiquement compositionnelles. C'est pourquoi une traduction directe serait inefficace.

(2) [spk1] et / puis / mes parents c'est les seuls à être partis / euh avec un oncle / un oncle euh / hm hm / qui a épousé donc euh une femme qui euh qui est du côté de Rouen / et donc euh son père si tu veux te- était grossiste euh en / hm hm /
[spk1] en comment dirais-je en en produits de vaisselle
[spk1] je sais pas comment on appelle ça enfin [ESLO2 ENT_1022]

Cette définition large nous permet d'englober un large éventail de formules, comme nous le verrons plus loin.

2.2. Typologie des formules

2.2.1. Classifications dans les études phraséologiques et interactionnelles

Les PPI ont été étudiées dans le cadre de différentes approches, tant en phraséologie qu'en analyse interactionnelle. Plusieurs typologies phraséologiques ont déjà été développées pour rendre compte de ces « phrases toutes faites », même si, à notre connaissance, aucune description exhaustive sur corpus n'a encore été entreprise, en particulier pour le français. Cette question a été abordée par plusieurs auteurs en français. Par exemple, Fónagy (1982) distingue les « énoncés liés », spécifiques à certaines situations (salutations ou félicitations), des formules plus génériques, par exemple liées à l'expression d'un sentiment. Cowie (2001), différencie les « formules routinières », très contraintes du point de vue situationnel (et proches des *pragmatèmes* de Mel'čuk (2015) et Blanco & Mejri (2018)) des *formules de discours*, qui s'appliquent à des situations plus larges. Cependant, à notre connaissance, c'est dans la phraséologie espagnole et allemande que l'on trouve les typologies les plus avancées.

Dans la phraséologie espagnole, Zuluaga (1980) identifie un sous-ensemble d'énoncés phraséologiques fonctionnellement marqués : « 'dichos' et phrases toutes faites » dont l'interprétation se fait en contexte, clichés dialogaux, formules propres aux textes narratifs et formules figées pragmatiques, qui renvoient à des situations sociales stéréotypées, comparables aux *pragmatèmes* de Mel'čuk (2015) et Blanco & Mejri (2018). Un autre type concerne les énoncés phraséologiques interjectifs, à visée essentiellement expressive. Comme Zuluaga, Corpas Pastor (1996) distingue plusieurs types de formules routinières, mais c'est López Simó (2016) qui propose, à notre connaissance, la typologie la plus fine, avec 4 types principaux d'énoncés, subdivisés en plusieurs sous-types :

1) Les formules interpersonnelles, liées aux interactions sociales (*Bon rétablissement*, *Après vous*) ;
2) Les formules personnelles, qui ont souvent une fonction affective, émotionnelle (*Le pied !*, *On aura tout vu !*) ;
3) Les formules impersonnelles, non directement liées aux participants du dialogue mais plutôt sur des faits (*Ça se pourrait*) ;
4) Les formules méta-communicatives (*La séance est ouverte*, *à vrai dire*).

Dans la phraséologie allemande, on trouve également des typologies détaillées comme celle de Lüger (2007) qui propose deux groupes principaux d'expressions : les expressions ayant une fonction sociale large (contact, identité, etc.) et les expressions ayant une fonction discursive (fonctions évaluatives, communicatives, etc.).

En analyse conversationnelle, ces formules ont été théorisées comme des « routines conversationnelles » (Coulmas 1979, 1985). Les typologies qui ont été développées sont souvent basées sur la théorie des actes de langage (cf. « actes de langage stéréotypés » Kauffer (2019) ; voir aussi Ronan (2015)). Il existe diverses études sur des contextes spécifiques d'utilisation de la langue, comme les interactions commerciales (Kerbrat & Traverso 2008). Il existe également des études sur les modalisateurs du discours (Perrin 2012). À ce jour, peu d'études proposent des typologies générales basées sur des corpus.

Pour notre part, la typologie que nous avons choisie est proche de celle de López-Simó, mais nous verrons que de nombreuses formules appartiennent simultanément à plusieurs types.

2.2.2. Typologie adoptée

Classer les PPI de manière tranchée est une entreprise ambitieuse, étant donné que de nombreuses formules sont polysémiques et qu'une même formule peut avoir plusieurs fonctions.

Nous avons donc opté pour des classes non exclusives. Bien que certaines soient incompatibles, la plupart des catégories que nous avons identifiées peuvent être combinées. Nous utilisons les principales classes identifiées par Tutin (2019) :

a) Les formules méta-énonciatives : elles traitent du contexte d'énonciation et commentent ce qui est dit et la manière dont c'est dit.
1) la plupart d'entre elles sont des clauses parenthétiques.
2) elles sont facultatives et souvent déplaçables.

b) Les formules réactives : elles expriment une réaction – opinion, évaluation ou émotion – à l'interaction ou à une situation. Elles sont souvent clausatives, mais il existe aussi des utilisations parenthétiques (Pausé *et al.* 2022).

c) Les phrases situationnelles : énoncé interprétable uniquement en fonction du contexte.

d) Les formules rituelles (pragmatiques) : associées à des situations sociales ou de communication spécifiques et contraintes.

Chaque classe contient des types décrits dans le tableau 1. Comme nous le verrons, l'élément de base de notre système d'annotation est le type.

Classes	**Types**	**Description**
Meta-énonciative	Méta-linguistique (MMet)	Commentaire sur le contenu référentiel du message : reformulation, approximation, correction, etc. Exemple : o*n peut le dire comme ça*
	Méta-conative (MCon)	Appel de l'attention de l'interlocuteur sur le contenu du message. Exemple : t*u vois*
	Méta-négociative (MNeg)	Indication d'une négociation interne [une réserve] du locuteur sur son adhésion au contenu référentiel du message. Exemple : *je pense*
Réactive	Expressive (RExp)	Expression d'une émotion du locuteur au regard d'un objet ou d'une situation. Exemple : *et comment !*
	Intéractive (RInt)	Réaction du locuteur à ses propos ou ceux d'autrui, ou à un événement ou une situation. Exemple : *je suis d'accord*
Situationnelle	Evaluative (SEval)	Appréciation du locuteur d'une situation ou d'un objet Exemple : *pas mal !*
	Non évaluative (SitNEval)	Affirmation interprétable seulement au regard de la situation d'énonciation. Exemple : *ça me dit quelque chose*
Rituelle (pragmatèmes)	Salutations (PrS)	Formule rituelle spécifique aux salutations d'ouverture et de fermeture d'une interaction. Exemple : *comment vas-tu ?*
	Politesse (PrP)	Formule rituelle spécifique aux marques de politesse. Exemple : *prends ton temps.*
	Autre (PrA)	Autres formules rituelles. Exemple : *que puis-je vous offrir ?*

Tableau 1 : Types de PPI

Comme indiqué précédemment, une PPI peut avoir plusieurs types et la catégorisation dépend fortement du contexte d'énonciation. Dans l'exemple (3), *bien sûr* est utilisé pour spécifier la réponse positive du vendeur à la requête du client. Dans l'exemple (4), le locuteur indique une concession concernant l'évidence d'un fait[3].

(3) [*Dans une fromagerie, un client passe commande*]
C10 des fromages blancs faisselle de six
VE2 oui
VE2 <PPI type="RInt">bien sûr</PPI> [fromagerie]

(4) VE2 vous avez essayé de nouveaux établissements ou pas lyon
C17 euh oui on a fait le le celsius
VE2 ah <PPI type="SitNEval">ça me dit quelque chose</PPI> celsius
C17 c'était le nouveau là confluence
VE2 ah oui
VE2 <PPI type="MNeg">bien sûr</PPI>
VE2 dans l'ancien local de le bec
C17 oui oui [fromagerie]

Le tableau 2 présente des exemples de formules associées aux types fonctionnels.

PPI	**MMet**	**MCon**	**MNeg**	**RExp**	**RInt**	**SEval**	**SitNEval**	**PRS**	**PRP**	**PRA**
Qu'est-ce que je vous sers ?										+
Bien sûr ![1]					+					
Bien sûr[2]			+							
C'est pas grave						+			(+)	
Tu te rends pas compte		+								
Vas-y		+							(+)	
On peut le dire comme ça	+									
Comme vous disiez	+	+								
La honte !				+		+				
Bonne idée					+	+				

Tableau 2 : Exemple d'attribution de type à des PPI

[3] Corpus introduit en section 3.2.

3. Annotation des formules

3.1. État de l'art

L'annotation pragmatique est en plein essor et plusieurs projets ont émergé ces dernières années. Pour l'anglais, nous pouvons mentionner les projets *Dialogue Annotation and Research Tool* et *SPICE-Ireland Corpus* (Weisser 2019, Ronan 2015). Archer et al. (2008) présentent également des travaux basés sur des schémas d'actes de dialogue. Une étude française récente intitulée *Théorie de la langue en acte* est basée sur C-ORAL-ROM (Cresti et al 2011). Ces travaux ne sont pas centrés sur les formules : les éléments figés et libres sont annotés indistinctement en fonction des catégories d'actes de langage. Pour ce qui est de la phraséologie, Eshkol-Taravella & Grabar (2017) se sont intéressées à la typologie des marqueurs de reformulation avec une annotation dans ESLO et dans un corpus construit à partir du forum Doctissimo. Néanmoins, cette étude ne concerne qu'une sous-partie des formules préfabriquées. À notre connaissance, aucun travail à grande échelle n'a été entrepris sur les expressions figées pragmatiques.

3.2. 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) (Debaisieux & Benzitoun 2020). Il comprend un sous-ensemble représentatif de données textuelles écrites et orales librement disponibles. La partie orale constitue le corpus le plus important et le plus diversifié de ce type, comprenant de nombreux échantillons de français parlé : conversations, entretiens, réunions diverses et discours publics, pour un total de 3 088 443 mots. Le corpus ORFEO – la partie orale du corpus – comprend des annotations syntaxiques en dépendances (Kahane & Gerdes 2020, Deulofeu & Valli 2020, Nasr et al. 2020). Le corpus peut être téléchargé à partir de la plateforme Ortolang[4].

Pour ce travail d'annotation exploratoire, nous avons sélectionné des échantillons appartenant à différents genres d'interaction : conversation informelle, entretiens, interactions dans un magasin et réunion de travail (voir tableau 3). Tous les participants sont de langue maternelle française.

[4] https://www.ortolang.fr/market/corpora/cefc-orfeo (consulté le 02/03/2023).

Sous-corpus	Contexte	Locuteurs	Nombre de mots	Corpus source
repas_francais	Repas partagé par deux amies. Elles discutent de leurs études et des travaux qu'elles doivent rendre	2 femmes, étudiantes (22 ans)	8883	Clapi
commerce_ fromagerie	Interactions commerciales dans une fromagerie	3 vendeurs: 1 femme et 2 hommes (20-50 ans), plusieurs clients	23670	Clapi
reunion_ conception_ mosaic_ architecture	Réunion de travail entre architectes	1 femme architecte d'intérieur (30-40 ans) et 2 hommes architecte d'intérieur et architecte (30-40, 40-50 ans)	15229	Clapi
Isabelle_ Legrand_F_32_ Anne-Lies_Simo-Groen_F_30_RO	Entretien sociolinguistique sur la vie de quartier	3 femmes, 2 d'environ 30 ans et 1 d'environ 60 ans	16966	CFPP

Tableau 3 : Composition du corpus

3.3. Principaux éléments du schéma d'annotation

Dans le processus d'annotation des PPI, l'identification des formules a été réalisée de manière semi-automatique à l'aide d'un dictionnaire de données contenant une liste de 4211 phrases préfabriquées[5] et d'une grammaire élémentaire appliqués sur le corpus à l'aide de l'outil Nooj (Silberztein 2016)[6]. Ensuite, nous avons effectué une première sélection et délimité les formules retenues avec les balises <PPI></PPI>. Nous avons ensuite parcouru manuellement l'ensemble du corpus avec l'aide des versions audio pour étiqueter d'autres formules. Nous avons dû prendre des décisions concernant diverses problématiques telles que l'annotation des éléments discontinus, la délimitation des formules (devons-nous inclure les éléments syntaxiques dépendants ?) et les variantes lexicales.

3.3.1. Délimitation des PPI et des variantes

Les formules sont annotées avec l'élément <PPI> et l'attribut « type » avec un ou plusieurs des types présentés dans le tableau 2 *supra*. Les variantes ne sont pas discriminées pour le moment

[5] Dictionnaire élaboré à partir d'une liste de phrases préfabriquées relevées manuellement au cours des travaux d'élaboration de la classification sémantico-pragmatique.
[6] Nous utilisons une procédure incrémentale en enrichissant graduellement la liste des PPI intégrée à NooJ.

mais seront traitées par lemmatisation (cf. *infra*). Les modificateurs (soulignés dans les exemples suivants) sont intégrés dans la PPI lorsqu'ils sont insérés dans la formule (exemple 5) et non à la périphérie (exemple 6).

(5) <PPI type="RExp_SEval">c'est assez chouette</PPI> [fromagerie]

(6) oui <PPI type="RInt">je vois</PPI> très bien [entretien]

3.3.2. Éléments discontinus

En raison des règles de segmentation appliquées lors de l'élaboration du corpus (Nasr et al. 2020), certaines tournures sont divisées en plusieurs segments appartenant à des tours de paroles différents, comme par exemple (7). Nous utilisons *next* et *prev* pour indiquer le lien entre les parties d'une même formule, selon le modèle de la French Treebank (Abeillé *et al.* 2019).

(7) MAR <PPI_next type="SEval">il y a pas de</PPI>
MAR <PPI_prev>problèmes</PPI> [repas_francais]

3.3.3. Arguments facultatifs

Certaines PPI ont un argument optionnel, comme par exemple (8) dans lequel l'argument de *ça n'a rien à voir* n'est pas toujours exprimé. Nous avons introduit <VAL></VAL> pour indiquer le marqueur de valence active.

(8) a. c'est sûr que ça <PPI type="SEval">ça n'a rien à voir</PPI> [entretien]
b. <PPI type="SEval">ça n'a rien à voir</PPI> <VAL>avec</VAL> le lieu d'habitation [entretien]

Cette balise est également utilisée pour délimiter les marqueurs de valence des verbes de modalité épistémique (*cf.* rection faible ; Blanche-Benveniste & Willems 2007, Apothéloz 2003), comme par exemple (9).

(9) [...] <PPI type="MNeg">je pense</PPI> <VAL>que</VAL> après si tu si tu te donnes des délais il y a un moment où forcément tu vas te dire merde il faut que je me bouge [repas_francais]

3.3.4. PPI et constructions

Durant le processus d'annotation, nous identifions, en plus des formules, des constructions (Fillmore 2008, Croft & Cruse 2004, Lakoff 1987) spécifiques à l'oral. Nous les délimitons avec les balises <CONSTR></CONSTR> pour des études ultérieures. Ce sont des

combinaisons partiellement lexicalisées et des tournures spécifiques, sémantiquement compositionnelles. La balise <VAR> est utilisée pour délimiter les variables qui font partie de la valence active de la formule ou de la construction, comme dans l'exemple (10).

(10) <CONSTR>c'est quoi</CONSTR> <VAR>une Bagnolétaise<VAR> alors [entretien]

Cette première étape du schéma d'annotation nous permet d'identifier et de caractériser les formules à un niveau macro. La deuxième étape offrira une description sémantique plus précise.

3.4. Éléments secondaires du système d'annotation

3.4.1. Étiquettes sémantiques

La distinction faite selon les classes et les types introduits dans le tableau 1 n'est pas suffisante pour des applications didactiques ou linguistiques. Afin d'offrir une base de données de formules et d'exemples annotés utilisables à des fins pédagogiques, il est nécessaire d'ajouter des informations fonctionnelles et sémantiques comme le propose Bidaud (2002) dans son dictionnaire des *Structures figées de la conversation*.

Nous avons donc proposé des étiquettes pragmatico-sémantiques étroitement liées aux types fonctionnels, telles que :

- 'reformulation' (méta-énonciative > métalinguistique) : *on peut le dire autrement*, *je veux dire* ;
- 'accord' (réactif > interactif) : *c'est bon pour moi*, *c'est noté*, *on fait comme ça* ;
- 'ouverture de l'interaction' (pragmatème > ouverture) : *vous allez bien*, *ça fait longtemps*, *mes respects* ;
- 'expression de la surprise' (réactif > expressif) : *la vache*, *c'est dingue.*

Contrairement aux types fonctionnels, les étiquettes ne peuvent pas être combinées. Lorsqu'une PPI appartient à plusieurs types, chaque type doit avoir une étiquette correspondante. Dans les exemples (11) et (12), les PPI qui expriment la difficulté sont employés. On constate que c'est chaud est à la fois évaluatif et expressif. La formule contient un côté émotionnel avec l'expression de la contrariété. Les PPI *c'est chaud* et *c'est pas évident* partagent une étiquette « difficulté_ appréciation » mais le premier est également étiqueté « contrariété ».

(11) VE2 <PPI type="SEval_RExp" label="appreciation_difficulté_ contrariété">c'est chaud</PPI> parce qu'on est en train de se dire quon a pas le temps [fromagerie]

(12) JUD <PPI type="SEval" label="appreciation_difficulté">c'est pas évident</PPI> [repas_francais]

Notons que l'émotion exprimée par une PPI dépend fortement du contexte d'énonciation. C'est l'une des particularités des formules expressives qui sont souvent liées à un paradigme d'émotions. Ainsi, *c'est chaud* peut exprimer la contrariété, la tension ou l'excitation.

Cette liste d'étiquettes est encore en cours de développement et est élaborée de manière empirique, en partie sur la base de typologies existantes (entre autres, Bidaud 2002, López Simó 2016, Gharbi 2020).

3.4.2. Modalité

Notre schéma d'annotation n'inclut pas l'annotation des actes de langage selon la théorie d'Austin et Searle, contrairement à Ronan (2015) et Weisser (2019). Cependant, il nous semble important d'annoter la modalité de manière simple, afin de différencier plusieurs types de formules. Rappelons également que notre corpus ne comporte pas de ponctuation, conformément à la pratique habituelle d'annotation des corpus parlés, ce qui ne nous permet pas, par exemple, d'identifier les formules interrogatives avec le point d'interrogation.

Par commodité, nous n'utilisons que trois valeurs, basées principalement sur les marqueurs morphologiques et lexicaux et sur la prosodie :

a) L'assertion : elle comprend les phrases déclaratives, mais aussi les phrases exclamatives, car il est pratiquement impossible de distinguer ces deux derniers types sur la base d'indices formels.

b) L'interrogation : correspond aux clauses de type interrogatif, qu'elles comportent des marqueurs spécifiques tels que *qui* comme dans *c'est de la part de qui*, ou non (ex. *c'est pour aujourd'hui ou pour demain*), où la prosodie doit être exploitée.

c) L'injonction : correspond à peu près aux ordres et aux commandements, et se réalise principalement par le mode impératif (ex. *prends-en de la graine*) et quelques formes subjonctives (ex. *que je ne t'y reprenne pas*).

La modalité syntaxique est essentielle pour distinguer plusieurs valeurs de formules. Par exemple, la formule très fréquente *ça va* peut avoir plusieurs valeurs selon la modalité syntaxique :

- Modalité assertive : *c'est bon*, *c'est ok*, formule situationnelle ;
- Modalité assertive : *je vais bien*, formule interactionnelle réactive ;
- Modalité interrogative : *c'est bon ?*, formule réactive interactionnelle ;
- Modalité interrogative : *comment allez-vous ?*, formule rituelle.

3.4.3. Lemmatisation et questions relatives aux variantes

La lemmatisation des formules est essentielle pour rendre compte de la variabilité de ces éléments en fonction de plusieurs paramètres :

- Variation grammaticale nombre/personne : *comment vas-tu* vs *comment allez-vous ?*
- Variation de temps/mode : *il manquait plus que ça* vs *il manquerait plus que ça*
- Inclusion de modificateurs, comme mentionné *supra*.
- Omission du marqueur de négation ne : *ça ne fait rien* vs *ça fait rien*.

Toutefois, il convient de rappeler que cette variation n'est pas systématique pour chaque paramètre. Par exemple, la formule *tu parles* n'est pas attestée à la deuxième personne du pluriel (le *tu* a ici une valeur générique : il ne désigne pas l'interlocuteur).

Il est nécessaire de choisir une forme comme lemme parmi les variantes. Nous avons décidé de choisir les paramètres les plus courants pour le français parlé, en sélectionnant le présent, en favorisant la deuxième personne du singulier et la négation sans *ne*. Le tableau 4 illustre l'annotation des formules qui font partie du paradigme sémantique de l'évaluation de la difficulté.

PPI	Lemme	Modalité	Type	Etiquette	Format d'annotation
c'est chaud	c'est chaud	déclarative	SEval+RExp	appréciation difficulté + contrariété	<PPI lemma="c'est chaud" mod="declar" type="SEval_ RExp" label="appr_ diff_ contrar">c'est chaud</ PPI>
c'est super chaud	c'est chaud	déclarative	SEval+RExp	appréciation difficulté + contrariété	<PPI lemma="c'est chaud" mod="declar" type="SEval_ RExp" label="appr_ diff_contrar">c'est super chaud</PPI>
ça va être sport	ça va être sport	déclarative	SEval+RExp	appréciation difficulté + contrariété	<PPI lemma="ça va être sport" mod="declar" type="SEval_RExp" label="appr_ diff_ contrar">ça va être sport</PPI>
c'est pas évident	c'est pas évident	déclarative	SEval	appréciation difficulté	<PPI lemma="c'est pas évident" mod="declar" type="SEval_RExp" label="appr_ diff">c'est pas évident</PPI>

Tableau 4 : Illustration des paramètres d'annotation

4. Résultats de la première phase d'annotation

Au moment de la rédaction de cet article, 4 échantillons du corpus parlé ont été annotés par 2 annotateurs experts. Une première évaluation sur un échantillon du corpus (15222 mots) montre un accord inter-annotateur de 67% sur l'annotation des types dans cette première phase. Les désaccords ont été discutés et réglés de manière concertée.

Quelques observations intéressantes ont pu être faites à partir des corpus annotés, bien que les résultats soient à prendre avec précaution compte tenu de la petite taille de l'échantillon.

4.1. Productivité

Nous pouvons observer des différences intéressantes dans le nombre et la variété des PPI. Elles sont presque deux fois plus nombreuses dans la conversation informelle que dans la réunion de travail (voir Figure 1). La spontanéité des interactions, plus importante dans le contexte informel que dans le contexte professionnel, semble avoir un impact sur la productivité des formules. On constate également une productivité plus faible dans le contexte de l'entretien sociolinguistique qui correspond également à une interaction partiellement supervisée. La fréquence remarquable du contexte commercial peut s'expliquer par le grand nombre de locuteurs, par l'importance des formules rituelles (salutations, remerciements) et par le fait que les clients réguliers entretiennent des relations informelles avec les vendeurs.

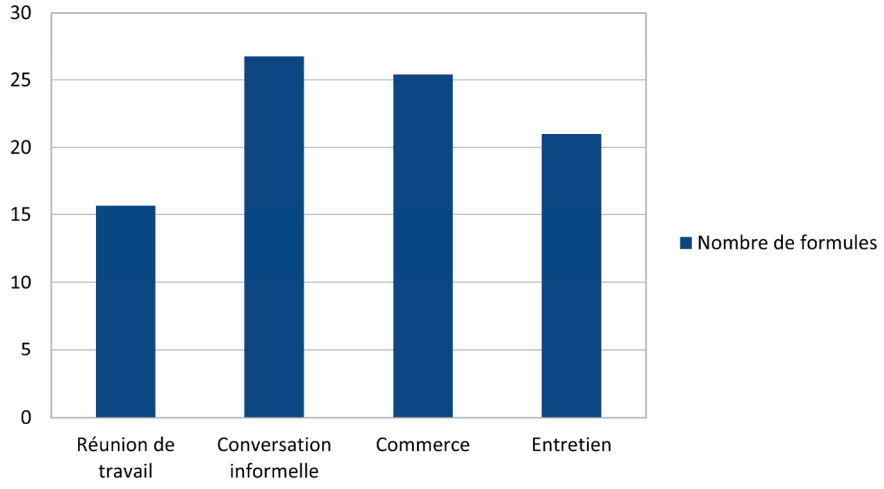

Figure 1 : Nombre de PPI pour 1000 mots

Si l'on s'intéresse maintenant à la variabilité des formules (rapport entre le nombre total de formules et le nombre de formules différentes), nous observons une plus grande variété dans les interactions amicales (2,9) et commerciales (2,9) que dans les interactions professionnelles (3,6) et les entretiens (3,9).

4.2. Distribution des types de PPI

L'identification des formules les plus fréquentes dans les 4 échantillons révèle que de nombreuses PPI ne semblent pas spécifiques à un sous-genre particulier. Par exemple, *c'est vrai* ou *d'accord* figurent parmi les 10 expressions les plus fréquentes dans les 4 sous-genres.

Cela n'est pas très surprenant, puisque ces formules se réfèrent à l'accord et à la confirmation, qui sont présents dans toutes sortes de textes. Il semble bien qu'il y ait des formules transgenres, comme l'observe Tutin (2019), même s'il faut rappeler que la plupart des formules sont polyfonctionnelles.

L'observation des types fonctionnels montre des spécificités intéressantes (comme on peut le voir sur la Figure 2) :

- La réunion de travail est ponctuée de formules de négociation et d'interaction visant à parvenir à une décision collective en évaluant les situations possibles. Les marqueurs de négociation tels que *c'est vrai*, *je trouve*, *je pense* sont particulièrement nombreux. Les conversations dans les magasins sont caractérisées par la présence, sans surprise, de salutations (*au revoir*, *bonne journée*, *à bientôt*) et par la forte présence de marques d'interaction directe liées aux transactions d'achat (*d'accord*, *c'est bon*, *bien sûr*).
- L'entretien sociolinguistique met en scène des locuteurs en situation asymétrique. Le discours de l'interviewé est proche d'un monologue où le locuteur cherche les formulations les plus appropriées, en utilisant des marqueurs méta-énonciatifs (*on va dire*) ou en essayant de modérer ses propos (*on va dire*). De l'autre côté, l'intervieweur utilise des marqueurs de feed-back comme *d'accord* pour encourager l'interlocuteur à poursuivre la discussion.
- La conversion informelle se caractérise par la forte présence de conatifs (*tu vois*, *tu sais*, etc.) et de marqueurs de négociation (*je pense*, *je crois*), liée à la proximité des locuteurs.

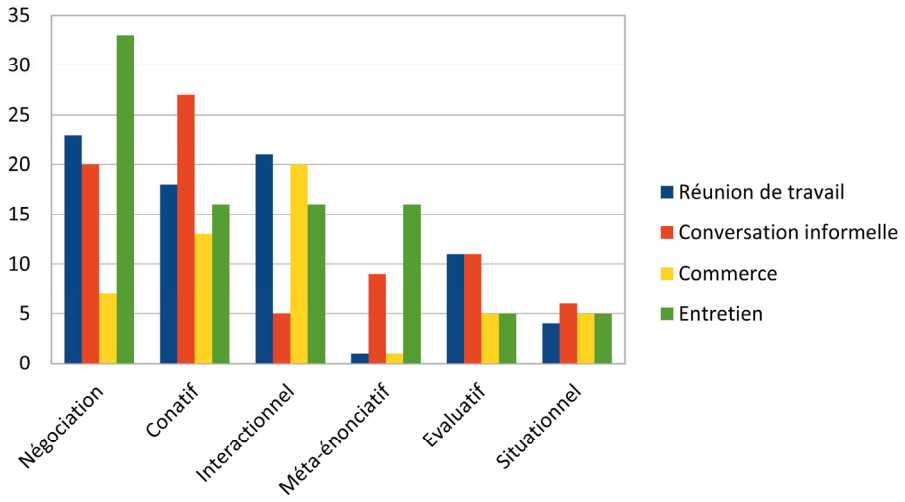

Figure 2 : Répartition des PPI en fonction des principaux types

L'étiquetage fonctionnel des PPI de l'interaction commerciale et de l'entretien montrent bien la plus grande variété des formules dans l'interaction commerciale. En effet, nous y avons attribué 76 étiquettes sémantico-pragmatiques à 201 formules contre 55 étiquettes pour 89 formules. Les Tableaux 5 et 6 montrent les 5 étiquettes les plus récurrentes pour chacun des types d'interaction.

Label	**Nombre de PPI**	**Exemple**
indication_validation	18	*c'est noté*
salutations_ouverture	14	*ça va ?*
appreciation_positive	10	*c'est bien parti*
appreciation_conditions_positives	9	*il y a pas de problème*
mrq_subj	8	*je me disais*

Tableau 5 : 5 étiquettes sémantico-pragmatiques les plus récurrentes dans l'interaction commerciale

Label	**Nombre de PPI**	**Exemple**
marqueur_subjectivité	7	*je dirais*
marqueur_recherche	5	*comment dire*
expression_résignation	4	*tant pis*
indication_accord	3	*tout à fait d'accord*
indication_accord_partiel	3	*pourquoi pas*

Tableau 6 : 5 étiquettes sémantico-pragmatiques les plus récurrentes dans l'entretien

5. Conclusion

Les formules préfabriquées interactionnelles sont un thème central de la phraséologie qui doit faire l'objet d'une attention particulière concernant leur traitement dans les corpus. Cependant, cette question est assez complexe, car elle requiert des compétences non seulement en sémantique et en syntaxe, mais aussi en pragmatique et en analyse interactionnelle. Le travail d'observation et d'annotation nécessite également une bonne connaissance des corpus oraux. Nous avons proposé ici une typologie de formules basée sur des classes et des types et validée en corpus par une annotation systématique. Les étiquettes sémantiques sont développées en parallèle avec la modalité et la lemmatisation.

Cette première étape d'annotation a permis la comparaison de formules dans différents genres discursifs, mettant alors en évidence, de manière prévisible, de grandes différences d'usage. Ces premières observations nous encouragent à développer le travail d'annotation sur un plus grand nombre d'échantillons.

Références bibliographiques

Abeillé, A., Clément, L., Liégeois, L. (2019), « Un corpus arboré pour le français : le French Treebank », *TAL*, 60/2, p. 19-43.

Apothéloz, D. (2003), « La rection dite "faible": grammaticalisation ou différentiel de grammaticité ? », *Verbum*, 25/3, p. 241-262.

Archer, D., Culpeper, J., Matthew D. (2008), "Pragmatic annotation", in Kytö, M., Lüdeling, A. (éds), *Corpus Linguistics: An International Handbook*, Mouton de Gruyter, Berlin, p. 613-42.

Bidaud, F. (2002), *Structures figées de la conversation. Analyse contrastive français-italien*, Peter Lang, Bern/Berlin.

Blanche-Benveniste, C., Willems, D. (2007), « Un nouveau regard sur les verbes faibles », *Bulletin de la Société de Linguistique de Paris*, 102, p. 217-254.

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

Corpas Pastor, G. (1996), *Manual de fraseología española,* Gredos, Madrid.

Coulmas, F. (1979), "On the sociolinguistic relevance of routine formulae", *Journal of pragmatics*, 3/3-4, p. 239-266.

Coulmas, F. (1985), *Conversational routine: Explorations in standardized communication situations and prepatterned speech*, Walter de Gruyter, Berlin.

Cowie, A. P. (ed.), (2001), P*hraseology. Theory, Analysis, and Applications,* Oxford University Press, Oxford.

Cresti, E., Massimo, M., Tucci, I. (2011), « Annotation de l'entretien d'Anita Musso selon la Théorie de la langue en acte », *Langue française*, 170, p. 95-110.

Croft, W., Cruse, A. (2004), *Cognitive linguistics*, Cambridge University Press, Cambridge.

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

Deulofeu, J., Valli, A. (2020), « Lexique et classement en parties du discours dans ORFÉO », *Langages*, 219, p. 53-68.
Eshkol-Taravella I., Grabar N. (2017), « Taxinomie dans les reformulations du point de vue de la linguistique de corpus », *Syntaxe et Sémantique*, 18, p. 149-184.
Fillmore, Ch. (2008), "Frame Semantics Meets Construction Grammar", in Bernal, E., De Cesaris, J. (eds) *Proceedings of the XIII EURALEX International Congress*, Institut Universitari de Lingüística Aplicada, Barcelone, p. 49-69.
Fónagy, I. (1982), *Situation et signification*, John Benjamins Publishing, Amsterdam.
Gharbi, N. (2020), *Analyse sémantico-pragmatique et discursive: les formules expressives de la conversation*. Thèse de doctorat, Université Grenoble Alpes-Université de Sfax.
Gosselin L. (2010), *Les modalités en français: la validation des représentations*, Rodopi, Netherlands/New York.
Kahane, S., Gerdes, K. (2020), « Annotation syntaxique du français parlé : Les choix d'ORFÉO », *Langages*, 219, p. 69-86.
Kauffer, M. (2013), « Le figement des "actes de langage stéréotypés" en français et en allemand », *Pratiques: théories, pratique, pédagogie*, 159-160, p. 42-54.
Kauffer, M., (2019) « Les "actes de langage stéréotypés" : essai de synthèse critique », *Cahiers de lexicologie*, 114/1, p. 149-171.
Kerbrat-Orecchioni, C., Traverso, V. (éds), (2008), *Les interactions en site commercial : invariants et variation,* ENS éditions, Lyon.
Krzyżanowska, A., Grossmann, F., Kwapisz-Osadnik, K. (2021), *Les formules expressives de la conversation Analyse contrastive : français-polonais-italien*, Episteme, Lublin.
Lakoff, G. (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago University Press, Chicago.
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.
Lüger, H. (2007), "Pragmatische Phraseme: Routineforme", in Burger, H. (ed.) *Phraseologie: ein internationales Handbuch der zeitgenössischen Forschung*, Mouton de Gruyter, Berlin, p. 444-459.
Martins-Baltar, M. (1995), « Énoncés de motif usuels : figures de phrase et procès en déraison », *Cahiers du français contemporain*, 2, p. 87-118.
Mel'čuk, I. (2016), "Clichés, an understudied subclass of phrasemes", *Yearbook of Phraseology*, 6/1, p. 55-86.
Nasr, A., Dary, F., Bechet, F., Fabre, B. (2020), « Annotation syntaxique automatique de la partie orale du ORFÉO », *Langages*, 219, p. 87-102.
Nuyts, J., Van der Auwera, J. (eds), (2016), *The Oxford Handbook of Modality and Mood*. Oxford University Press, Oxford.
Pausé, M. S., Tutin, A., Kraif, O., Coavoux, M. (2022), « Extraction de Phrases Préfabriquées des Interactions à partir d'un corpus arboré du français parlé : une étude exploratoire », in Neveu, F., Prévost, S., Steuckardt, A., Bergounioux, G., Hamma, B. (éds), *8e Congrès Mondial de Linguistique Française (CMLF) 2022,* SHS Web of Conferences, 138, Orléans.

Perrin, L. (2012), « Modalisateurs, connecteurs, et autres formules énonciatives », *Arts et Savoirs*, 2, http://journals.openedition.org/aes/500 (consulté le 14/04/2022).
Ronan, P. (2015), "Categorizing expressive speech acts in the pragmatically annotated SPICE Ireland corpus", *ICAME Journal*, 39, p. 25-45.
Silberztein, M. (2016), "Formalizing Natural Languages: the NooJ Approach", Hoboken, NJ USA.
Tutin, A. (2019), « Phrases préfabriquées des interactions : quelques observations sur le corpus CLAPI », *Cahiers de Lexicologie*, 114/1, p. 63-91.
Weisser, M. (2019), "The DART annotation scheme: form, appliability & application", *Studia Neophilologica*, 91/2, p. 131-153.
Zuluaga, A. (1980), *Introducción al estudio de las expresiones fijas,* Peter Lang, Berne.

Corpus

CLAPI : http://clapi.ish-lyon.cnrs.fr/ [consulté le 01/04/23]
CFPP : http://cfpp2000.univ-paris3.fr/ [consulté le 01/04/23]