Marqueurs intonosyntaxiques en français parlé et genres :

Compter pourquoi, compter quoi, compter comment ?

Julie Beliao, Anne Lacheret, Sylvain Kahane

Laboratoire MoDyCo, UMR-7114, Université Paris-Ouest, France
julie@beliao.fr, anne@lacheret.com, sylvain@kahane.fr

1 Introduction

La typologie textuelle fait l'objet d'une solide tradition en linguistique de l'écrit (Jakobson 1963, Bakhtine 1984), en particulier dans le cadre de l'analyse de discours (Swales 1990, Maingueneau 1996, Adam 1999, entre autres), et reste bien ancrée dans les recherches actuelles sur les corpus écrits qui font massivement appel à la notion de fréquence (B. Habert *et al.* 1997). Elle est, en revanche, encore balbutiante en ce qui concerne l'oral : si elle existe, c'est essentiellement dans une perspective comparative avec l'écrit et si nous avons accès à des sources de référence pour l'anglais (Halliday 1989, Biber *et al.* 2000), nous n'en disposons pas pour le français.

Quand la fréquence est mobilisée à l'oral, elle est explorée essentiellement dans une perspective variationniste[1] : dans la continuité des recherches de Labov, c'est la fréquence sociale qui mobilise principalement les chercheurs, i.e. la recherche de corrélations entre traits formels et propriétés socio-démographiques. L'objectif prioritaire est d'analyser, selon un angle d'attaque essentiellement lexical, morphosyntaxique ou phonématique les corrélations entre les profils variables de locuteurs, la variation des usages sociaux et les marques qu'ils laissent dans le texte. Quant à la perspective envisagée sur les comptages, elle est le plus généralement unitaire, autrement dit, quand bien même, l'analyse repose sur une série d'unités et que ces dernières peuvent être représentées par des fréquences variables de traits, elles sont toujours envisagées indépendamment les unes des autres dans l'analyse et non corrélativement à leurs différentes combinaisons temporelles dans la chaîne parlée.

Dans ce contexte, notre contribution présente les deux caractéristiques originales suivantes : (i) étendre le champ de la typologie textuelle au domaine de l'oral, *i.e.* se focaliser sur la fréquence textuelle pour laisser de côté la fréquence sociale ; (ii) explorer pour ce faire les marqueurs intonosyntaxiques des textes, indépendamment d'abord, de façon combinée ensuite. Le second point a nécessité le développement d'un cadre méthodologique innovant pour l'annotation de ces deux couches, qui permette à la fois de les explorer de façon modulaire et autonome, et d'explorer leurs corrélations. Étant donné ce cadre d'analyse intonosyntaxique, prendre comme angle d'attaque la fréquence textuelle ne signifie pas, contrairement à la pratique courante, compter le nombre d'occurrences d'un trait formel par unités de texte (à l'écrit, des mots) ; il devient crucial de tenir compte − et c'est là où réside l'essentiel de la complexité de la tâche − de la dimension temporelle de la chaîne parlée.

En pratique, ce type d'approche consiste à poser l'hypothèse suivante : les variations d'occurrence de traits prosodiques et syntaxiques, et leurs différents types de combinaisons dans les textes, rendent compte de compétences communicatives variables auxquelles sont associés différents genres de discours. Autrement dit, une situation de communication donnée, caractérisable par des variables situationnelles plurielles[2] engendre un genre discursif particulier, doté de marqueurs formels spécifiques, caractérisables en termes de traits langagiers prototypiques (ce sont nos *variables descriptives*) et identifiables en tant que tels dans les textes.

La tâche est alors la suivante : définir les variables situationnelles et les variables descriptives dont l'étude des corrélations va permettre de caractériser et de classer des productions orales, et rendre compte de la triple association :

Situation de communication → Marqueurs formels $_{\{syntaxiques/prosodiques/intonosyntaxiques\}}$ → Type de genre

En linguistique de corpus, cela suppose de (i) disposer d'un matériel échantillonné selon des variables situationnelles dont le choix doit être argumenté en fonction d'hypothèses précises, annoté en syntaxe et en prosodie pour faire émerger les corrélations, (ii) extraire les variables descriptives des corpus à corréler avec les variables situationnelles, là encore selon une démarche argumentée quant au choix des variables, (iii) mettre en place une méthode d'annotation qui

[1] Que la variation soit diastratique, diatopique ou stylistique. A cet égard, les travaux conduits dans le cadre du projet *Phonologie du français contemporain* sont particulièrement représentatifs (Detey *et al.* 2010).

[2] Les variables situationnelles sont les traits consignés dans les métadonnées en fonction de la situation de communication. Pour illustration, une situation dialogale peut être interactive, semi-interactive (cf. *infra*, § 1.1, Table 2).

ne perde pas les informations temporelles dans l'encodage des unités syntaxiques, (iv) conduire une méthode statistique dont les choix sont clairement explicités et justifiés : de même qu'il n'existe pas de corpus omnibus, il n'existe pas de méthode statistique tout-venant, et celle-ci doit être adaptée au type de données manipulées.

Cet article présente la problématique à travers l'exploitation du treebank Rhapsodie (Lacheret *et al.* à par. 2015a), un corpus constitué de 57 échantillons relativement courts de français parlé (5 minutes en moyenne), soit 3 heures de parole[3] (33000 mots, 89 locuteurs) munies d'une transcription orthographique (segmentation en mots) et phonétique (segmentation en phonèmes, syllabes et pauses)[4], et richement annotées. La section 2 présente le corpus design ainsi que les annotations syntaxiques et prosodiques réalisées sur les données. La section 3 répond aux questions : *compter quoi ?* et *compter comment ?* D'une part, elle a pour objectif de présenter la sélection des variables descriptives utilisées pour réaliser nos comptages et conduire nos corrélations. Par ailleurs, il s'agit de justifier le choix des méthodes statistiques retenues pour explorer nos données, vu les caractéristiques atypiques du corpus Rhapsodie, en particulier l'hétérogénéité des données et la durée variable des échantillons, et d'en présenter les principes essentiels. Enfin, nous illustrons par des exemples variés, les premiers résultats obtenus quant à la caractérisation et à la classification des échantillons Rhapsodie. Ce dernier volet apporte une illustration concrète sur l'apport de différentes méthodes d'apprentissage automatique pour la modélisation des genres de discours.

2 Corpus

Dans cette section, nous exposons dans un premier temps la constitution du réservoir Rhapsodie, *i.e.* les choix effectués pour l'échantillonnage et la constitution des métadonnées (variables situationnelles retenues pour l'analyse). Les principes d'annotation syntaxique et prosodique sont ensuite présentés.

2.1 Corpus design

Au cœur du projet Rhapsodie : (i) l'objectif de modéliser l'interface intonosyntaxique sur un jeu de constructions, annotées en prosodie et en syntaxe, suffisamment vaste pour permettre les généralisations descriptives, (ii) l'hypothèse selon laquelle il existe une relation étroite entre les caractéristiques typologiques d'un texte, *i.e.* les patrons textuels définis sur les bases de critères strictement formels et le genre de discours dont il est issu, *i.e.* les traits situationnels qui le caractérisent[5]. En conséquence, les lignes directrices pour l'échantillonnage du corpus Rhapsodie ont été les suivantes : (i) collecter un ensemble d'échantillons suffisamment diversifié en termes de typologie textuelle, (ii) disposer d'un panel de locuteurs assez large pour éviter les idiosyncrasies individuelles, (iii) étant donné ces deux premières contraintes et étant donné le coût temporel colossal que suppose une annotation robuste[6] sur le versant de la syntaxe comme de la prosodie, les échantillons sont nécessairement courts (cinq minutes en moyenne).
Dans la mesure où il n'existe pas à l'heure actuelle de corpus de référence pour le français parlé dans lequel nous aurions pu aller puiser pour construire notre réservoir et atteindre cet objectif de diversité et d'équilibre typologique, nous avons, dans un premier temps extrait nos données de sources existantes (sources institutionnelles, dont notamment PFC (Durand *et. al* 2009), CFPP 2000 (Branca-Rosof *et. al* 2012) et ESLO (Eskhol-Taravella *et. al* 2012))[7]. Ce premier jeu d'échantillons a ensuite été complété par différents types de données (multimédia, descriptions de films, descriptions d'itinéraires entre autres) récoltées dans le cadre du projet Rhapsodie afin d'assurer l'équilibre de l'échantillonnage fixé au préalable.
Si les linguistes s'accordent pour considérer qu'un genre de discours peut être décrit comme un objet multifactoriel qui intègre un ensemble de variables socio-communicatives a priori orthogonales (Biber *et al.* 1999, Koch & Oesterreicher 2001)[8] et si, en conséquence, les angles

---

3 Réalisé dans le cadre de l'ANR Rhapsodie 07 Corp-030-01, http://www.projet-rhapsodie.fr.

4 Les transcriptions alignées au son ont été effectuées sous PRAAT sur différentes tires. Les chevauchements de paroles (partiels ou complets) sont consignés dans la tire orthographique. La transcription phonétique, sur deux tires (segmentation en phonèmes et en syllabes), a été réalisée semi-automatiquement avec le logiciel Easy-align (Goldman 2011).

5 Voir aussi le concept de « registre » chez Biber & Conrad (2009), caractérisé par une série de traits lexicaux-grammaticaux récurrents dans les textes d'une certaine variété et qui servent des fonctions communicatives majeures.

6 Par *annotation robuste*, nous voulons dire des annotations nettoyées et vérifiées pour celles qui reposent sur l'utilisation d'outils automatiques, des annotations fondées sur des campagnes inter-annotateur pour celles qui ont été réalisées manuellement.

7 La description exhaustive des sources est accessible sur le site Rhapsodie http://www.projet-rhapsodie.fr/propriete-intellectuelle.html.

8 Un genre de discours peut être décrit selon la nature de la situation de communication (localisation des interlocuteurs, but communicationnel, degré de formalité), mais aussi selon l'environnement spatial et le canal de communication, ou encore en fonction du contenu thématique, etc.

d'attaque pour caractériser un genre sont pluriels, nous avons privilégié cinq traits situationnels majeurs. En premier lieu, en distinguant les monologues des dialogues, nous opposons les discours produits par un unique locuteur à l'intention d'une large audience ou d'un petit auditoire et les discours produits par au moins deux locuteurs, plus ou moins interactifs. Ensuite, l'opposition parole privée vs. parole publique distingue les échantillons tirés d'entretiens en face à face qui peuvent être extraits d'interactions au quotidien ou avoir été réalisés en présence d'un informateur, et les conférences ou les émissions TV et radiophoniques. Dans les deux types de parole, les thèmes de discours sont larges et diversifiés, dans la parole publique, les émissions de nature variée (entretiens politiques, débats d'idées, émissions de vulgarisation scientifique, etc.). Au final, chaque type de parole (monologale vs dialogale), qu'il soit privé ou publique est renseigné selon les variables suivantes : taux de planification, degré d'interactivité, type de séquence (Table 1)[9].

[9] Les métadonnées sont renseignées selon le schéma IMDI sous le logiciel Arbil (http://tla.mpi.nl/tools/tla-tools/arbil/) développé au Max Planck à Nijmegen (Withers 2012).

Type de parole	**Privée/publique**	**Monologue**
		Dialogue
	Planification (spontané, semi-spontané, planifié)	
	Interactivité (interactif, semi-interactif, non-interactif))	
	Séquence discursive (argumentative, descriptive, procédurale, oratoire)	

Table 1 Variables situationnelles dans Rhapsodie

2.2 Annotations macrosyntaxiques

Le système d'annotation syntaxique développé pour le projet Rhapsodie comprend principalement un étiquetage morpho-syntaxique, une structure de dépendance fonctionnelle, et un découpage en unités macrosyntaxiques, que nous appelons *unités illocutoires*. C'est sur ce dernier que nous avons travaillé dans le cadre de cette étude.
Le niveau macrosyntaxique (Berrendonner 1990, Cresti 2000, Benzitoun *et al.* 2010) définit la cohésion illocutoire à l'intérieur de l'énoncé. Cette structure est composée d'une unité centrale appelée « noyau », qui concentre la force illocutoire de l'énoncé, et éventuellement d'unités satellites définies selon des critères strictement topologiques. Contrairement au noyau, ces unités satellites ne sont pas autonomes du point de vue illocutoire et forment avec lui un seul et même énoncé. On peut voir ci-dessous différentes configurations « satellites » possibles de l'unité illocutoire : l'exemple (1) est une construction en « pré-noyau – noyau », le (2) en « pré-noyau – noyau – post-noyau », le (3) en « noyau – post-noyau » et le (4) amorcé par un introducteur *mais*[10].

(1) pour eux < c'est important // [Rhap-D0006, CFPP 2000]
(2) déjà < j'ai retrouvé mes origines > quand même // [Rhap-D1003, Rhapsodie]
(3) qu'est-ce que vous en pensez > de la boule magique // [Rhap-D2011, Rhapsodie]
(4) ^mais non // [Rhap-D0004, CFPP 2000]

Les modèles macrosyntaxiques existants peuvent être stratificationnels ou modulaires, *i.e.* considérer les unités maximales de la microsyntaxe comme des entrées pour le niveau macrosyntaxique ou au contraire envisager des traitements indépendants. Le modèle défini dans le cadre du projet Rhapsodie se différencie des modèles stratificationnels comme, par exemple, celui proposé par Berrendonner (1990) qui considère les unités maximales de la microsyntaxe, comme des points d'ancrage pour la macrosyntaxe. La macrosyntaxe de Rhapsodie s'apparente plutôt aux modèles modulaires comme ceux élaborés par l'école d'Aix-en-Provence (Blanche-Benveniste 1990) ou par Cresti (2000), qui considèrent les deux types d'organisation comme orthogonaux et donc comme pouvant opérer de concert sur les mêmes séquences.
En outre, le modèle de Rhapsodie pose l'hypothèse que l'organisation macrosyntaxique répond à un principe de cohésion qui opère de manière indépendante de la prosodie. Au final, deux contraintes principales ont été retenues pour segmenter un énoncé en unités illocutoires (désormais UI) ; (i) caractériser l'organisation syntaxique indépendamment de toute organisation prosodique ou du moins de toute théorie prosodique ; (ii) proposer des critères syntaxiques de segmentation explicites permettant aux annotateurs d'appliquer de manière aussi formelle que possible les choix théoriques présidant à l'annotation du corpus. Le principal critère retenu est la non-autonomie : les segments qui ne peuvent former un énoncé autonome[11] sont considérés comme dépendants macrosyntaxiquement.
Précisons qu'une UI peut contenir une ou plusieurs autres UI selon deux modalités : (i) par enchâssements d'UI (le discours rapporté en (5) et les greffes[12] en (6)) notés [ ] ou (ii) par insertions d'UI (les parenthèses en (7)) notées ( ).

(5) Marcel Achard écrivait [elle est très jolie // elle est même belle // elle est élégante //] // [Rhap-D2001, Corpus Mertens]
(6) vous suivez la ligne du tram qui passe vers la [je crois que c'est une ancienne caserne //] // [Rhap-M0003, Corpus Avanzi (Avanzi 2012)]
(7) alors que Heinze (c'est quand même assez extraordinaire hein //) c'est le patron de la défense // [Rhap-D2003, Rhapsodie]

[10] La fin des pré-noyaux est indiquée par le symbole <, le début des post-noyaux par > et l'unité illocutoire se termine par //.
[11] C'est-à-dire interprétable en tant que tel par l'interlocuteur.
[12] Nous appelons greffe une UI qui vient occuper un position régie où est attendue en principe une unité lexicale : « *cette fille, c'est une pousse toi de là que je m'y mette* ». Les discours rapportés sont également des UI qui viennent occuper une position régie, mais à la différence des greffes, il s'agit d'une position où est attendue une telle construction.

Enfin, nous désignons par *noyau associé* (voir (8)), des segments munis d'un opérateur illocutoire, mais qui ne véhiculent pas de contenu informationnel. Ces unités sont équivalentes à celles définies par Morel et Danon-Boileau (1998) sous le terme de *marqueurs discursifs*, terminologie que nous adopterons ici.

(8) « ouais » il y a un accident « quoi » // [Rhap-M0023, Rhapsodie]

2.3 Annotations prosodiques

L'annotation prosodique est effectuée en deux étapes, une étape manuelle (Avanzi *et al.* 2015) qui consiste à coder les proéminences et les disfluences syllabiques perçues, une étape automatique qui génère les différentes unités de la structure prosodique (Lacheret *et al.* à par. 2015b).

L'annotation manuelle a été réalisée par cinq codeurs naïfs, un taux d'accord inter-annotateurs a ensuite été calculé pour dériver une annotation intermédiaire vérifiée et stabilisée par trois experts pour générer l'annotation de référence[13]. En pratique une proéminence syllabique est une syllabe proéminente qui se détache de son environnement phonétique (Terken 1991), une unité pourvue d'une ou de plusieurs syllabes proéminentes est une unité qui se dégage comme une figure sur un fond discursif. Une échelle à 3 degrés a été retenue pour le codage des proéminences : syllabe non proéminente ('0'), syllabe modérément proéminente ('W'), syllabe fortement proéminente ('S')[14]. Le concept de disfluence (prosodique) est communément utilisé pour désigner des points dans la chaîne parlée correspondant à une perturbation du programme syntagmatique (Blanche-Benveniste & Jeanjean 1986). Il s'agit d'une classe générique qui regroupe les pauses remplies (ou pauses d'hésitation) précédées et/ou suivies de 'euh', les allongements syllabiques supérieurs au seuil d'allongement qui marque une syllabe accentuée, les répétitions, les faux départs et les inachèvements de morphèmes, de mots ou de syntagmes. Ces phénomènes apparaissent souvent de façon combinée dans le flux discursif et sont marqués, du moins nous en posons l'hypothèse, par des patrons temporels et mélodiques spécifiques.

L'annotation dérivée automatiquement de l'annotation manuelle génère une structure hiérarchique composée de quatre niveaux de constituance avec du plus bas au plus haut de la hiérarchie le *pied métrique*, le *groupe rythmique*, le *paquet intonatif* et la *période intonative*[15]. Si les trois premiers et leurs différents types sont annotés sur les bases de proéminences et des disfluences perçues, la dernière, celle qui nous intéresse ici, est segmentée semi-automatiquement avec le logiciel Analor[16] sur les bases de la méthode développée par Lacheret & Victorri (2002) pour la segmentation de données monologales. Dans un flux de parole produit par un et un seul locuteur, la frontière potentielle d'une période, est détectée si et seulement si les quatre conditions suivantes sont remplies: 1) présence d'une pause d'au moins 300 ms, 2) détection d'un mouvement de F0 qui atteint une certaine amplitude, définie comme la différence de hauteur entre le dernier extremum et la moyenne de F0 sur toute la portion du signal précédant la pause, 3) détection d'un saut, défini comme étant la différence de hauteur entre le dernier extremum de F0 précédant la pause et la première valeur de F0 suivant la pause, 4) absence de « euh » dans le voisinage immédiat de la pause. Il convient de souligner que la décision de reconnaissance d'une rupture périodique repose sur un principe de compensation de seuils En d'autres termes, la détection ne dépend pas des valeurs exactes des paramètres, mais de leurs seuils respectifs d'activation et des poids associés[17] : quand un paramètre est légèrement au-dessous du seuil choisi ('-1'), une frontière de période est détectée si les autres paramètres ont des valeurs clairement au-dessus du seuil (ex. Figure 1).

---

[13] La perception des proéminences et des disfluences est un phénomène complexe qui comporte une certaine part de subjectivité. Ainsi pour les proéminences, si l'auditeur s'appuie sur des indices acoustiques pour leur identification, il ne peut y avoir de perception brute indépendamment des contraintes grammaticales qui déterminent en partie les attentes des auditeurs. Une campagne inter-annotateurs supervisée par un contrôle d'experts s'avère donc indispensable (Lacheret *et al.* 2010).

[14] Où les marqueurs 'W' et 'S' correspondent aux valeurs *weak* et *strong*.

[15] Pour une présentation détaillée des différents niveaux de constituance, voir Lacheret *et al.* (2011).

[16] http://www.lattice.cnrs.fr/analor.

[17] Activation très forte : '2', forte : '1', moyenne : '0', en dessous du seuil : '-1'.

Figure 1: Segmentation en 3 périodes de l'énoncé : [je pense aux nombreuses victimes de la tempête $_{PI}$ et à toute leur famille endeuillée $_{PI}$ dont nous partageons la peine $_{PI}$] [D2004, Corpus Rhapsodie]. Les deux lignes verticales représentent les frontières des trois périodes. Avec pour la première période (respectivement les seuils 2, -1, 1 ; pour la seconde : 1, 1, 2 ; pour la dernière : 2, 1, 2).

Dans le cadre du projet Rhapsodie, il a fallu étendre la segmentation en périodes aux fichiers de dialogues et donc tenir compte des contextes de chevauchements de parole[18]. Les chevauchements peuvent donner lieu à trois types de périodes intonatives : (i) les périodes tronquées à droite (l'interlocuteur essaye de prendre le tour de parole sur le locuteur courant puis abandonne laissant le locuteur courant fini), pour lesquelles, les caractéristiques acoustiques de la fin de la période du locuteur en cours sont masquées étant donné l'interruption de son interlocuteur, (ii) les périodes tronquées à gauche pour lesquelles, ce sont les caractéristiques acoustiques du début de la période du locuteur en cours qui sont masquées, (iii) les périodes tronquées à gauche et à droite (combinaison des deux cas précédents). Soit respectivement les exemples suivants extraits des échantillons [Rhap-D0009, PFC], exemples (1) et (2) et [Rhap-D0001, CFPP2000] :

Figure 2: Illustration de trois types de périodes tronquées, en (1) la période est tronquée à droite, en (2) la période est tronquée à gauche et en (3) la période est tronquée à droite et à gauche.

3 Analyse des données

Pour mémoire, l'analyse du corpus repose sur l'exploration combinée des variables situationnelles, fournies par les métadonnées, et des données primaires et secondaires. Nous désignons par *données primaires* le signal brut ainsi que les différentes transcriptions, orthographique et phonétique, qui incluent les chevauchements de parole (*overlaps*). Les annotations prosodiques et syntaxiques constituent les données secondaires. Beaucoup de propriétés formelles sont potentiellement observables et peuvent donc devenir des variables statistiques pertinentes pour estimer la fréquence des constructions ; d'où quatre questions : (i) y a-t-il des traits plus discriminants que d'autres pour caractériser et classer nos échantillons ? (ii) parmi l'ensemble des traits observables, certains sont-ils redondants ? (iii) dans quelle mesure les données secondaires constituent un apport significatif pour ces opérations de caractérisation et de classification ? (iv) cet apport, s'il est constaté, est-il général ou dépend-t-il des variables situationnelles et/ou formelles manipulées ? Nous n'avançons aucune réponse a priori, seule l'étude quantitative, qui fait l'objet de cette section, nous permettra de

[18] Quand les tours s'enchaînent sans chevauchement, il y a segmentation automatique en périodes, qu'ils soient ou non séparés par une pause et ce quelles que soient les configurations acoustiques en jeu.

répondre.

3.1 Compter quoi ?

3.1.1 Traits retenus pour l'analyse des données primaires

Les comptages sur les données primaires s'appuient simultanément sur (i) des variables langagières brutes qui peuvent être calculées indépendamment des annotations syntaxiques et prosodiques (ii) des transcriptions orthographique et phonétique. Ces éléments présentent l'avantage de ne mettre en jeu aucun a priori théorique sous l'angle intonosyntaxique et de constituer de ce point de vue un espace « neutre » pour notre analyse en fréquences[19]. Comparer les résultats obtenus des données primaires et secondaires traitées séparément, puis conjointement nous permettra de diagnostiquer l'apport de l'annotation pour la tâche de classification des échantillons de Rhapsodie. De manière plus profonde, de tels résultats renseignent sur la façon dont l'approche Rhapsodie, à l'interface de la prosodie et de la syntaxe, peut contribuer aux recherches en typologie textuelle sur l'oral.
Pour les données primaires, le focus est mis sur le calcul des variables descriptives qui sont en fait des fréquences, l'ensemble des fréquences calculées dans cette étude sont des fréquences temporelles représentent le nombre de fois qu'un phénomène linguistique se reproduit par unité de mesure du temps, ici la seconde. Ce choix distingue les fréquences que nous manipulons des fréquences textuelles habituellement utilisées par les linguistes, il repose sur l'hyptohèse que, pour l'étude d'un corpus de langue parlée, l'unité de mesure temporelle est incontournable et supplante le mot.
On calcule donc des fréquences temporelles de tokens20 (fT)[21] – qui correspondent à un calcul de débit –, de pauses (fP) et d'overlaps (fO) (voir Table 2)[22]. Pour calculer ces fréquences, nous avons besoin de recueillir pour l'ensemble du corpus et pour chaque échantillon indépendamment : la durée de l'échantillon (dE), le temps de pause (tP) et le temps d'overlaps (tO), à partir desquels on calcule le temps d'articulation (tA) – durée de l'échantillon moins temps de pauses et d'overlaps[23] –, le temps de parole (tP) – durée de l'échantillon moins temps de pause[24] ; pour ce faire, le nombre de tokens par échantillon (nT), le nombre de pauses par échantillon (nP) et le nombre d'overlaps par échantillon (nO) sont préalablement calculés.
Ces informations et l'ensemble des traitements nécessaires à cette étude sont obtenus automatiquement grâce à la plate-forme *OOPS* (Beliao & Liutkus 2014) et sont stockées dans un tableau de données (Table 2).

3.1.2 Traits retenus pour l'analyse des données secondaires

Trois questions sous-tendent le traitement des données secondaires présentées ci-dessous : (i) comment évaluer le rendement descriptif, en termes de caractérisation des échantillons, de chaque type de variable manipulée, syntaxique, prosodique et intonosyntaxique ? (ii) que peut-on dire des caractéristiques complémentaires ou redondantes de ces variables[25] ? (iii) dans quelle mesure le traitement intonosyntaxique, i.e. la combinaison des variables prosodiques et syntaxiques, apporte de nouvelles informations par rapport à un traitement modulaire ? En particulier, comment les résultats fournis par l'analyse des relations entre unités illocutoires et périodes intonatives en termes de synchronisation et d'inclusion renseigne sur les découpages des textes en différents types d'unités élémentaires et, ce, de manière variable en fonction des types de discours en jeu ?
Nous allons répondre à ces questions en considérant séparément d'abord les annotations syntaxiques et prosodiques, puis en les considérant simultanément.

---

19 Nous ne prétendons pas que la transcription orthographique tout comme la segmentation phonétique ne reposent pas sur des choix théoriques (ex. la segmentation syllabique d'un texte peut varier sensiblement en fonction du modèle phonologique sous-jacent). Il s'agit juste de dire que ces choix n'ont pas d'incidence directe sur le type d'étude qui nous occupe ici.

20 On entend ici par token : mot orthographique.

21 L'ensemble des variables utilisées est regroupé par catégories et par niveau de description dans la Table 2.

22 Étant donné notre programme de travail à l'interface de la syntaxe et de la prosodie, il ne nous a pas semblé nécessaire pour les unités segmentales d'extraire des unités d'un grain plus fin que le token (comme la syllabe et le phonème). Des connaissances sur le comportement de ces unités auraient été en revanche nécessaires dans une analyse phonostylistique.

23 Qui correspond à l'intervalle temporel sur lequel l'annotation prosodique n'a pu être réalisée que partiellement voire pas du tout étant donné la présence de chevauchements (cf. supra Figure 1).

24 L'unité de mesure temporelle est la seconde.

25 Dans le cas des variables complémentaires, chacune des variables contribue à la caractérisation de l'échantillon, dans le cas des variables redondantes, les deux types de variables apportent la même information, une seule suffit donc à la caractérisation.

- **Annotations syntaxiques**

Nous avons vu dans la section 2.2 qu'il existait différents types d'UI (± insérée, ± enchâssée, ± régie). Dans un premier temps, il semble raisonnable de ne pas les mettre sur le même plan au niveau des analyses statistiques ; ce, pour deux raisons : (i) se pose la question de savoir si le type d'UI et leur nombre pourrait avoir une influence sur la catégorisation des échantillons selon nos différentes variables situationnelles ; (ii) une réponse à cette première question nous permettra d'orienter la manière de construire nos variables intonosyntaxiques. Autrement dit, dans le but de ne pas assimiler des unités qui caractériseraient différemment les échantillons, sept variables ont été calculées à partir des comptages syntaxiques, ce sont les fréquences syntaxiques d'UI ordinaires (fUI), d'UI enchâssées (fUI_{ench}, voir exemples (5) et (6) section 2.2), d'UI insérées (fUI_{ins} , voir exemple (7) section 2.2), d'UI inférieures à trois tokens (fUI_{small} , voir exemple (4) section 2.2). Les deux dernières variables sont : la fréquence des marqueurs discursifs (fDM, voir exemple (8)), des tokens *euh* (f_{euh}) et de tokens *mh* (f_{mh})[26].

- **Annotations prosodiques**

Pour ce qui est des données prosodiques nous avons choisi de ne retenir que cinq variables. Le premier jeu de variables concerne la fréquence temporelle des différents types de proéminences qui ont été annotées dans le corpus : proéminences fortes (fPROM), proéminences faibles ($fPROM_W$), syllabes non proéminentes ($fPROM_{nul}$) et les proéminences en général – combinaison des proéminences fortes et faibles (fPROM) ; le second jeu est relatif aux périodes intonatives (fPI) ; comme pour la syntaxe, cette étape à fait l'objet de choix, notamment en ce qui concerne les PI tronquées, que nous avons exclues. La fréquence de chaque unité prosodique est donnée par le ratio du comptage des unités et du temps d'articulation.

- **Combinaisons intonosyntaxiques**

Il s'agit ici d'obtenir de façon aussi précise que possible des informations quantitatives sur l'interface intonosyntaxique. Nous avons choisis de nous concentrer sur quatre variables différentes et a priori complémentaires.
La première variable, le rapport entre la fréquence d'UI et la fréquence de PI (fUI/fPI), est donnée en log afin de rendre la variable symétrique et ainsi de ne pas privilégier le taux UI par PI par rapport au taux de PI par UI[27]. Cette information indique potentiellement une relation d'inclusion entre ces deux types d'unités[28]. Les trois autres variables calculées nous renseignent sur les propriétés de synchronisations des UI[29] et des PI. Nous avons pour la présente étude répertorié trois types majeurs de synchronisation (voir Figure 2) : la synchronisation totale (UI et PI sont strictement équivalentes, elles partagent les mêmes frontière temporelles, (1)), la synchronisation droite (UI qui partagent la même frontière droite qu'une PI, (2)) et la synchronisation gauche (UI qui partagent la même frontière gauche qu'une PI, (3)). Il existe d'autres configurations, atypiques, telles qu'elles sont présentées en (4) et (5), et que nous ne comptons pas directement[30]. En conséquence les variables suivantes ont été retenues : le ratio fUI/fPI, la fréquence des UI totalement synchronisées avec les PI ($f_{UI\equiv PI}$, celle des UI et PI synchronisées à droite ($f_{UID\equiv PID}$) et celle des UI et PI synchronisées à gauche ($f_{UIG\equiv PIG}$).

---

[26] Le statut des « mh » en tant qu'unité illocutoire n'ayant pas été tranché lors de l'annotation, cette dernière reste hétérogène sur ce point.

[27] En effet, log(a/b) = -log(b/a).

[28] Un ratio fUI/fPI supérieur à 1 nous indique que l'échantillon contient plus d'UI que de PI, l'UI sera donc probablement l'unité qui inclut des PI.

[29] Étant donné que les overlaps n'ont pas reçu d'annotations en PI, nous avons considéré pour ce calcul uniquement les UI hors overlaps.

[30] Ces configurations sont indirectement comptabilisées par les différences entre les trois valeurs considérées.

Figure 3 : Types de synchronisation possibles des frontières syntaxiques et prosodique, où (1) représente une synchronisation totale, où bornes gauches et droites d'UI et de PI coïncident temporellement (il s'agit d'un cas particulier qui ne peut se combiner avec aucun autre), (2) représente une synchronisation droite, (3) une synchronisation gauche, (4) une non-synchronisation par chevauchement et (5) une non-synchronisation par inclusion.

Variables	Glose	Unité de mesure	Catégorie de données	Niveau de description
fT	Fréquence de tokens	par seconde	Primaire	
fP	Fréquence des pauses	par seconde	Primaire	
fO	Fréquence d'overlaps	par seconde	Primaire	
dE	Durée de l'échantillon	en secondes	Primaire	
tP	Temps de pause	en secondes	Primaire	
tO	Temps d'overlapping	en secondes	Primaire	
tA	Temps d'articulation	en secondes	Primaire	
tPl	Temps de parole	en secondes	Primaire	
nT	Nombre de tokens	en tokens	Primaire	
nP	Nombre de pauses	en unité de pauses	Primaire	
nO	Nombre d'overlaps	en unité d'overlaps	Primaire	
fUI	Fréquence des UI	par seconde	Secondaire	Syntaxe
fUI_{ench}	Fréquence des UI enchassées	par seconde	Secondaire	Syntaxe
fUI_{ins}	Fréquence des UI insérées (incises)	par seconde	Secondaire	Syntaxe
fUi_{small}	Fréquence des « petites » UI (< 3 mots)	par seconde	Secondaire	Syntaxe
f_{DM}	Fréquence de marqueurs discursifs (d*iscursive markers*)	par seconde	Secondaire	Syntaxe
f_{euh}	Fréquence des *euh*	par seconde	Secondaire	Syntaxe
f_{mh}	Fréquence de *mh*	par seconde	Secondaire	Syntaxe
$fPROM_{nul}$	Fréquence des syllabes non proéminentes	par seconde	Secondaire	Prosodie
$fPROM_{W}$	Fréquence des syllabes proéminentes faibles (*weak*)	par seconde	Secondaire	Prosodie
$fPROM_{S}$	Fréquence des syllabes proéminentes fortes (*strong*)	par seconde	Secondaire	Prosodie
fPROM	Fréquence des syllabes proéminentes faibles et fortes	par seconde	Secondaire	Prosodie
fPI	Fréquence des périodes intonatives	par seconde	Secondaire	Prosodie
UI/PI	Taux de UI par rapport à PI	taux	Secondaire	Intonosyntaxe
$f_{UI\equiv PI}$	Fréquence d'UI et PI partageant une frontière droite et une frontière gauche	par seconde	Secondaire	Intonosyntaxe
$f_{UID\equiv PID}$	Fréquence d'UI et PI partageant une frontière droite	par seconde	Secondaire	Intonosyntaxe
$f_{UIG\equiv PIG}$	Fréquence d'UI et PI partageant une frontière gauche	par seconde	Secondaire	Intonosyntaxe

Table 2. *Variables descriptives primaires et secondaires.*

3.2 Compter comment ?

On s'interroge ici sur la manière dont les données secondaires viennent enrichir et compléter la caractérisation des échantillons d'un corpus Les deux jeux de données, primaires et secondaires, ont été testés avec deux méthodes différentes de classification, les arbres de décision et les machines à vecteurs supports (SVM). Ces méthodes font clairement ressortir l'apport des données secondaires par rapport aux données primaires. Dans un deuxième temps, nous verrons comment l'étude conjointe de certaines variables secondaires, via une analyse en composantes principales, permet d'épingler des mécanismes redondants pour la classification.

3.2.1 Apprentissage supervisé : critères de caractérisation et de classification

Une fois les données récoltées, nous disposons pour chaque échantillon du corpus d'un vecteur de traits décrits plus

haut (variables descriptives), ainsi que des variables situationnelles rattachées à cet échantillon. L'objectif de cette section est de montrer que certaines techniques d'apprentissage automatique permettent de prédire convenablement le genre rattaché à un échantillon à partir de la seule observation en fonction de la fréquence des traits.
Le but de l'apprentissage est de construire un modèle général à partir de données particulières afin de prédire un comportement face à une nouvelle donnée. On cherche donc à répartir un ensemble d'éléments en plusieurs classes (variables situationnelles) en fonction de leurs traits (variables descriptives). Les classes peuvent être inconnues (cas de la classification non-supervisée) ou bien connues a priori (cas de la classification supervisée) comme dans notre étude. Une méthode d'apprentissage automatique fonctionne alors de la manière suivante : dans un premier temps, on lui fournit un ensemble de vecteurs de traits ainsi que la variable situationnelle qu'elle doit caractériser, par exemple le caractère planifié ou non du discours. L'algorithme d'apprentissage, spécifique au modèle utilisé, apprend alors sur cette base une manière de prédire la variable situationnelle à partir de l'observation du vecteur de traits. Lorsqu'on disposera ensuite d'échantillons dont on ne connaît pas le caractère planifié ou non, on pourra utiliser le modèle appris pour le prédire.
Dans le but d'évaluer les performances d'une telle technique d'apprentissage automatique, on procède souvent par validation croisée, c'est-à-dire que, dans notre cas, une partie seulement des 57 échantillons du corpus, appelée *base d'apprentissage*, sert à l'apprentissage du modèle de classification, tandis que l'autre partie, appelée *base de test*, est utilisée pour évaluer ce modèle. L'intérêt de procéder de la sorte est d'évaluer non seulement la capacité du modèle à expliquer des données connues et observées, mais également sa capacité à s'appliquer correctement à des données inconnues qu'il s'agit de classifier automatiquement, c'est à dire sa capacité de généralisation.
Dans cette étude, nous considérons deux modèles de classification automatique : les arbres de décision et les machines à vecteur support.

- **Arbres de décision**

Un arbre de décision prend en entrée un vecteur de traits et donne en sortie la valeur de la variable situationnelle étudiée, par exemple le caractère planifié du discours. La construction est descendante, c'est-à-dire qu'au début tous les échantillons sont regroupés. L'algorithme va alors diviser récursivement et le plus efficacement possible les échantillons de l'ensemble d'apprentissage par des tests définis à l'aide des variables jusqu'à ce que l'on obtienne des sous-ensembles d'échantillons appartenant à une même classe. Les arbres de décision ont trois qualités qui nous paraissent particulièrement appréciables pour le linguiste[31] : (i) la classification est très rapide, (ii) la décision y est réalisée de manière dichotomique, i.e. binaire, (iii) en conséquence, les décisions sont aisément interprétables (voir figure 4).
Les premiers algorithmes de classification par arbres de décision sont anciens. L'arbre est construit[32] de sorte que chaque embranchement conduise au meilleur partitionnement possible des données. En pratique, un partitionnement sera jugé bon s'il sépare les échantillons en deux lots qui ont chacun la même valeur pour la variable explicative (c'est-à-dire sans bruit), ou du moins dont la variance pour cette variable est la plus réduite possible.
La *validation croisée* désigne l'opération qui permet de tester la précision prédictive du modèle dans un échantillon test (parfois aussi appelé *échantillon de validation croisée*) par rapport à la précision prédictive de l'échantillon d'apprentissage à partir duquel le modèle a été développé. Cela revient à apprendre le modèle sur une partie du corpus utilisé et à vérifier ensuite le modèle sur le reste du corpus. Nous avons constaté que l'algorithme CART (Breiman 1984) utilisé présentait un taux d'erreur non négligeable en situation de *validation croisée*. Ceci s'explique par la grande sensibilité de l'algorithme à ce que l'on appelle communément « la malédiction de la dimension » (*curse of dimensionality*). Plus le nombre de variables descriptives (vingt-sept dans notre cas) et de variables situationnelles (deux-cent seize combinaisons potentielles à partir de 14 variables situationnelles) augmente, moins ce type d'algorithme est performant en raison d'une tendance au sur-ajustement. Autrement dit, il n'est pas difficile de construire un arbre de décision qui ne fasse aucune erreur sur un jeu de données. Cependant, il est fort probable qu'un tel arbre ait de très mauvaises capacités de généralisation. Pourquoi? Essentiellement parce que les tests qu'il effectue sont beaucoup trop spécifiques aux données d'apprentissage et ne capturent pas suffisamment les véritables manières de classifier la population dont sont issues les données de test.

- **SVM**

Les machines à vecteurs supports (SVM en anglais pour *support vector machine*) adoptent une approche différente des arbres de décision pour apprendre des classifieurs sur les données. Les SVM sont une généralisation de ce cas de figure simple au cas où les échantillons ne sont pas simplement envisagés selon le plan (abscisse, ordonnée) – et donc caractérisés par un vecteur – , mais évoluent dans un espace de dimension supérieure.
Sans rentrer dans les spécifications techniques, les machines à vecteur support procèdent à une séparation optimale en

31 Habitué à manipuler des arbres.
32 Nous avons utilisé le module « *rpart* » avec le programme de statistique R pour générer nos arbres.

deux ensembles (nous travaillons avec les deux composantes principales, mais il est possible d'en avoir plus) en se concentrant sur les quelques échantillons qui peuvent poser problème, parce qu'ils sont à la frontière des deux groupes. Cette particularité des SVM de ne considérer qu'un faible nombre d'échantillons pour procéder au traitement les rend particulièrement attractives pour le traitement de données que l'on peut caractériser par un grand nombre de traits. Ainsi contrairement aux arbres de décision qui cherchent le meilleur critère à chaque fois, les SVM cherchent la meilleure combinaison de critères.

- **Résultats et discussion**

Nous avons appliqué les deux méthodes de classification décrites plus haut avec une validation croisée de type *leave-one-out*, c'est-à-dire que le modèle est appris sur 56 échantillons du corpus et testé sur le dernier, en moyennant sur les 57 possibilités offertes par une telle approche, le tout de manière itérative. De manière générale, la technique *leave-one-out* permet d'estimer correctement l'erreur, avec un biais faible, mais une variance[33] plus élevée que les autres techniques.
Nous avons calculé le taux de réussite des méthodes, tantôt en ne donnant que les données primaires et tantôt en ne donnant que les données secondaires. On trouvera les résultats correspondants dans la table 3 pour les arbres de décision, et dans la table 4 pour les SVM.

[33] Le biais reste présent puisqu'on apprend le modèle sur la majorité des échantillons du corpus, étant donné la petite taille du corpus ; en conséquence la variance peut-être élevée sur l'échantillon qui peut toujours correspondre à un genre sous-représenté dans les échantillons d'apprentissage.

Taux de réussite (arbres)	Variables situationnelles				
	Type de parole (2 valeurs)	**Situation de communication (2 valeurs)**	**Planification (3 valeurs)**	**Interactivité (3 valeurs)**	**Type de séquence (4 valeurs)**
Données primaires (3 variables)	**86.4**	58.5	**66.4**	**60.5**	39.3
Données secondaires (17 variables)	74.2	**65.4**	54.7	46.7	39.3
Données primaire + secondaires	77.5	**65.4**	56.4	**60.5**	**57.1**

Table 3 Tableau récapitulatif des taux de catégorisation correcte des **arbres de décision** (en pourcentages) obtenus pour chaque classe en fonction des différents jeux de données.

Taux de réussite (SVM)	Variables situationnelles				
	Type de parole (2 valeurs)	**Situation de communication (2 valeurs)**	**Planification (3 valeurs)**	**Interactivité (3 valeurs)**	**Type de séquence (4 valeurs)**
Données primaires (3 variables)	82.45	66.66	56.14	66.66	36.84
Données secondaires (17 variables)	77.19	**85.96**	**64.91**	**68.42**	64.91
Données primaire + secondaires	**84.21**	82.45	59.64	64.91	**66.66**

Table 4 Tableau récapitulatif des taux de catégorisation correcte des **SVM** (en pourcentages) obtenus pour chaque classe en fonction des différents jeux de données.

De la table 3, il ressort que les performances des arbres de décision en termes de généralisation sont souvent plus mauvaises (type de parole, planning) ou assez peu améliorées dès lors qu'on utilise plus de traits pour classifier les échantillons. Cela peut s'interpréter en invoquant la « malédiction de la dimension », à laquelle les arbres de décision sont très sensibles. En substance, si on augmente le nombre de traits à considérer alors que l'on dispose de peu d'échantillons, le volume de l'espace correspondant augmente et les algorithmes d'apprentissages d'arbres de décision souffrent alors du faible nombre de données d'apprentissage. Pour le comprendre, imaginons qu'on ait une portion du plan à décrire, ou bien une portion de volume. On sent bien qu'il y aura besoin de plus de points pour décrire le volume que pour décrire la surface. À plus forte raison, si on considère des éléments qui ont une vingtaine de dimensions, il sera très difficile à des algorithmes gloutons comme ceux utilisés pour apprendre des arbres de décision de parvenir à une bonne solution globale.
Au contraire, les SVM présentent de meilleures performances de généralisation, même, et surtout, lorsque le nombre de traits augmente. Cela est dû au fait que l'algorithme d'apprentissage correspondant a la particularité de déterminer un optimum global au problème de la classification, tandis que celui des arbres procède de manière gloutonne (un partitionnement après l'autre), et au final une solution optimale dans l'immédiat, ne représentera pas forcement la solution optimale à long terme. Plusieurs propriétés des SVM expliquent leurs bonnes performances pratiques. Tout d'abord, une normalisation automatique des données selon chacun des traits est effectuée lors de la phase d'apprentissage. Cela permet de prendre en compte automatiquement la dynamique observée sur le corpus. Ensuite, le paramètre de marge permet de déterminer plus ou moins de vecteurs supports pour apprendre correctement le modèle. Il est significatif (eu égard au nombre d'échantillons représentatifs de chaque variable considérée) que dans nos tests, sur 56 échantillons utilisés pour l'apprentissage, 51 aient été utilisés comme vecteurs support. En effet, étant donné le faible peuplement de l'espace des traits par les échantillons d'apprentissage, il est normal que presque toutes les observations aient été conservées, puisque les données ne sont pas redondantes. Les arbres de décision n'offrent pas une telle souplesse.

Dans notre étude, quoique les performances obtenues par les arbres de décision soient honorables (autour de 60 % en moyenne), nous constatons un taux de réussite plus haut pour les SVM, autour de 70 %, et, ce, pour chaque variable situationnelle. La plupart du temps, le gain induit par la prise en compte des variables secondaires est notable. On observe notamment un meilleur score pour la situation de communication, la planification et l'interactivité. En revanche, on peut noter que pour le type de parole et de séquence, le score est meilleur en tenant compte conjointement des variables secondaires et primaires.

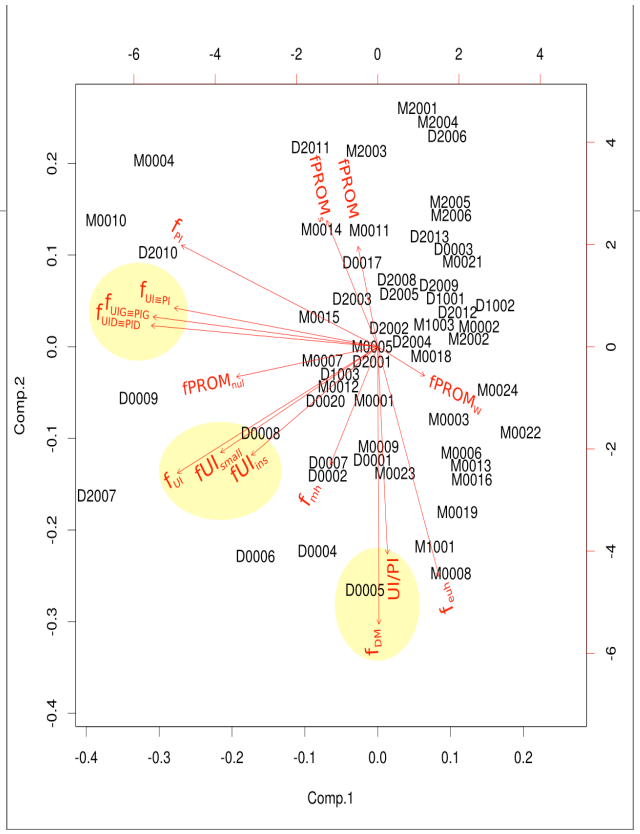
3.2.2 L'analyse en composantes principales

La conclusion que l'on tire des expériences précédentes est que le grand nombre de traits descriptifs pris en compte pour la classification selon les variables situationnelles étudiées engendre une difficulté pour les algorithmes d'apprentissage automatique. À cette problématique s'en ajoute une autre que nous allons à présent considérer : celle de la représentation de telles données multidimensionnelles. L'objectif de la tâche que nous présentons ici est de « résumer » nos données de dimension importante dans des espaces de faible dimension (deux dimensions : abscisses et ordonnées), à des fins de visualisation graphique et d'interprétation fonctionnelle.
Considérons les échantillons de notre corpus ainsi que chacun de leurs traits descriptifs, nous savons comment analyser séparément chacune des vingt-sept variables descriptives, soit en faisant un graphique, soit en calculant des résumés numériques ; nous savons également que l'on peut regarder les relations entre deux variables (par exemple fréquence des tokens et fréquence des overlaps), soit en faisant un graphique du type nuage de points, soit en calculant leur coefficient de corrélation linéaire, voire en réalisant la régression de l'une sur l'autre. Mais comment faire une représentation graphique simultanée des vingt-sept variables ? La difficulté vient de ce que les échantillons ne sont plus représentés dans un espace à deux dimensions, mais dans un espace de dimension vingt-sept. C'est l'objet de l'Analyse en Composantes Principales (ACP, Pearson 1901, Hotelling 1933, Jolliffe 2002) de représenter les données dans un espace de dimension réduite (ici : deux) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent de nos données. Autrement dit, l'ACP est une méthode qui consiste à transformer les variables descriptives, *a priori* liées entre elles, en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables, d'un nombre réduit, sont alors nommées « composantes principales », ou « axes principaux », elle présentent l'avantage de rendre compte de l'information de manière orthogonale *a priori*.

- **La redondance des traits**

Comment détecter les variables descriptives qui sont redondantes les unes par rapport aux autres, autrement dit en termes statistiques, qui co-varient[34].
La figure 5 illustre la redondance de certaines des variables secondaires choisies. Chaque groupe de variables redondantes a été mis en valeur par des pastilles. Ainsi, les variables f_{UI}, fUI_{small} et fUI_{ins} caractérisent tous les échantillons d'une manière similaire. Ce résultat est étonnant dans la mesure où l'on s'attendait à ce que les UI non-standard (enchâssées, insérées et courtes) caractérisent différemment les échantillons. En conséquence on peut en déduire que les UI peuvent être regroupées ensembles et négliger cette hypothèse.

[34] Dans la théorie des probabilités et statistiques, les concepts mathématiques de covariance et de corrélation sont très similaires. Les deux décrivent la mesure dans laquelle deux variables aléatoires ou des ensembles de variables aléatoires ont tendance à s'écarter de leurs valeurs attendues de façon similaire.



gure 4 ACP représentant tous les échantillons du corpus en fonction des variables secondaires. Les pastilles grisées ettent en évidence les variables redondantes, c'est-à-dire les variables caractérisant les échantillons selon un biais lativement proche. La proximité des vecteurs est à prendre en compte mais aussi leur longueur. Les vecteurs courts diquent des vecteur qui sont éloignés de l'hyperplan (= composante principale) et donc potentiellement assez différents.

- **La synchronisation/inclusion**

La figure 4 nous permet également de mettre en évidence l'importance de l'interface intonosyntaxique dans la caractérisation des échantillons. Comme présenté en section 3.1.2, nous avons retenu quatre variables intonosyntaxiques qui sont, pour mémoire, le ratio UI/PI, la fréquence des UI totalement synchronisées avec les PI ($f_{UI\equiv PI}$), celle des UI et PI synchronisées à droite ($f_{UID\equiv PID}$) et celle des UI et PI synchronisées à gauche ($f_{UIG\equiv PIG}$).
Nous remarquons sur la figure que la variable « UI/PI » fonctionne de concert avec la variable « marqueurs discursifs » (pastille basse), ce qui nous indique que les échantillons qui comptent un plus grand nombre d'UI que de PI, donc qui ont tendance à inclure des UI dans des PI, sont également caractérisés par une forte fréquence de marqueurs discursifs. On distingue dès lors deux groupes d'échantillons, les échantillons incluant des PI dans des unités syntaxiques et les échantillons incluant des UI dans des unités prosodiques. Dans une précédente étude (Beliao et Lacheret 2013), nous avons montré que la croissance de la fréquence des marqueurs discursifs était corrélée à la planification du discours : autrement dit, plus on compte de marqueurs du discours, plus le discours est improvisé. Ces résultats viennent se combiner avec le constat que les échantillons guidés par la prosodie (i.e. qui ont une forte tendance à inclure plusieurs UI dans une PI) sont plutôt spontanés, contrairement à ceux qui seraient guidés par la syntaxe (i.e qui, à l'inverse, ont tendance à inclure plusieurs PI dans une UI) et qui voient leur fréquence de marqueurs discursifs drastiquement chuter.
Autre fait notable, les deux variables de synchronisation semblent fonctionner ensemble (pastille en haut à gauche), c'est-à-dire qu'un échantillon qui présente beaucoup de synchronisation droite aura de grandes chances de présenter également beaucoup de synchronisation gauche. On remarque de plus que les échantillons ayant un grand nombre de synchronisations (droite, gauche ou totale) sont marqués dans la mesure où il sont minoritaires (environ une quinzaine d'échantillons) par rapport aux échantillons présentant un taux plus faible de synchronisations.
En comparant le ratio et les différents types de synchronisation, on constate que le taux d'UI par PI (pastille basse), qui nous informe sur le taux d'inclusion des UI dans les PI et vice versa, fonctionne orthogonalement (c'est à dire de manière complémentaire et non redondante) aux variables de synchronisation. En d'autres termes ce taux nous informe

sur la nature de l'unité la plus représentée (PI ou UI). Par exemple, les échantillons appartenant à la classe « oratoire » sont des échantillons qui ont tendance à être relativement synchronisés et à avoir un taux UI/PI inférieur à 1, donc à inclure des PI dans des UI (M2001, M2003, M2004, D2006), ces résultats confirment les résultats observés dans Beliao (2013). L'oratoire serait donc un type de séquence synchronisant surtout les UI avec les PI et pas l'inverse (si plusieurs PI étant incluses dans une UI, leurs frontières sont peu congruentes avec celle d'une UI) et de ce fait guidé par une programmation plutôt syntaxique, du moins nous en faisons l'hypothèse.
Ces affirmations sont toutefois à tempérer dans la mesure où l'ACP a été effectuée ici selon les deux premières composantes principales seulement (abscisse et ordonnée de l'hyperplan). Bien qu'on voit effectivement que selon ces deux composantes, les traits décrits ci-dessus se comportent essentiellement de la même manière, il est tout à fait possible qu'ils se distinguent plutôt dans des composantes tertiaires, qui permettent de caractériser les données avec plus de granularité.

4 Conclusions

La question des relations entre genres de discours et la fréquence textuelle, vue comme récurrence de propriétés formelles dans les textes, est féconde et abondamment diffusée dans le domaine de l'écrit, mais reste une question largement ouverte et peu explorée à l'oral, voire inexistante pour ce qui est de l'utilisation des fréquences intonosyntaxiques. Cette situation résulte principalement de la rareté des données disponibles. Parmi les raisons plurielles qui peuvent expliquer cette carence, nous invoquerons la complexité de la tâche (pas de modèles de référence ni de standards d'annotation, contraintes instrumentales lourdes et chronophages pour gérer l'alignement temporel des différentes couches d'annotation, étape complexe et incontournable dans ce type d'étude). Cette difficulté concerne aussi bien la production des données que leur analyse. (i) Quel type d'annotation proposer pour rendre compte non seulement de la présence ou de l'absence d'un objet quelconque, mais également de ses caractéristiques temporelles (où se situe-t-il dans la chaîne parlée et combien de temps dure-t-il), enfin de sa fréquence d'occurrence dans cet empan temporel ? Répondre à cette seconde question suppose qu'on ait recours à une méthode stable de segmentation en unités discrètes d'un objet par essence continu en ce qui concerne la couche d'annotation prosodique ? (ii) Comment produire des annotations syntaxiques et prosodiques alignées temporellement qui permettent ensuite de les interroger de façon combinée ? (iii) quelle architecture de données développer et quel langage choisir pour construire un système de requête optimal pour renseigner sur les fréquences de construction intonosyntaxiques dans les textes?
Concernant le dernier point, nous avons exploité une architecture de données objet pour aborder cette thématique dans le cadre du corpus Rhapsodie (Beliao & Liutkus 2014) en tenant compte de la spécificité de ce réservoir qui nous a demandé une extrême vigilance quant à l'utilisation des méthodes quantitatives classiques. En particulier : (i) la petite taille du corpus au regard des masses de données énormes traitées à l'écrit, (ii) l'hétérogénéité des sources, (iii) le déséquilibre dans les échantillons pour représenter certaines variables situationnelles (ex. très peu d'échantillons sur le versant des séquences oratoires) ont posé des questions cruciales en termes de normalisation des données d'une part et du choix des méthodes statistiques d'autre part[35]. C'est ces deux premières questions qui ont guidé nos choix statistiques afin de remplir efficacement la feuille de route que nous nous sommes fixés dans le cadre de ce numéro thématique sur la fréquence en linguistique, et de répondre à un jeu de quatre questions majeures que nous rappelons pour mémoire. (i) Dans quelle mesure la mise au jour de propriétés formelles (ou variables descriptives) syntaxiques, prosodiques et intonosyntaxiques, et d'informations sur leur fréquence d'occurrence, peuvent être corrélées à certaines variables situationnelles pour caractériser des types de textes et les genres qui les engendrent ? (ii) Quel est l'apport significatif des données secondaires, qui dérivent d'un processus d'annotation long et coûteux par rapport à de simples données primaires, pour la caractérisation des échantillons et leur classification en genres (iii) Parmi l'ensemble des propriétés explorées, comment se répartissent-ils en termes de redondance de l'information, de complémentarité ou d'indépendance des variables ? (iv) Dans quelle mesure des méthodes d'apprentissage automatique bottom-up (fréquences mesurées) permettent une approche réflexive approfondie de la notion de « fréquence» telle qu'elle est manipulée intuitivement par le linguiste ? Autrement dit, comment permettent-elles de falsifier mais aussi d'enrichir par de nouveaux descripteurs, sinon les théories du genre encore inexistantes à l'oral, du moins les hypothèses sur les corrélations possibles entre faits de syntaxe et faits de prosodie et conduire à des généralisations descriptives? Il ressort ainsi clairement de notre étude que certains traits syntaxiques, prosodiques et intonosyntaxiques constituent de bons prédicteurs des variables situationnelles. En particulier, nous avons vu que les échantillons ayant l'étiquette « spontané » sont caractérisés par un plus grand nombre de disfluences prosodiques, mais aussi par une plus grande fréquence de marqueurs discursifs. Des indices intonosyntaxiques viennent aussi nous renseigner, dans la mesure où l'on observe que le ratio UI/PI est caractéristique du genre de discours produit. Les échantillons comportant plus d'UI que de PI et dont les UI sont significativement plus courtes que les PI sont plutôt « non-planifiés », alors que ceux rassemblant les critères inverses appartiennent clairement à la catégorie « planifié », associée aux séquences oratoires.

[35] Cette difficulté en entrée d'analyse ne nous a pas empêché de trouver des descripteurs efficace pour la reconnaissance de ce genre.

Bibliographie

ADAM J.-M. (1999), *Linguistique textuelle : des genres de discours au texte*, Paris, Nathan.

AVANZI M., BORDAL G., LACHERET A., OBIN N., SAUVAGE-VINCENT J. (à par. 2015), *The annotation of syllabic proéminences and disfluencies* , in Lacheret-Dujour *et al.* Eds.

AVANZI M. (2012), « *L'interface prosodie/syntaxe en français Dislocations, incises et asyndètes »*, Bruxelles, Peter Lang.

BAKHTINE M. (1984), *Esthétique de la création verbale*, Paris, Gallimard.

BELIAO J., LIUTKUS A., (2014), *OOPS: une approche orientée objet pour l'interrogation et l'analyse linguistique de l'interface prosodie/syntaxe/discours*, In SHS Web of Conferences (Vol. 8, pp. 2565-2581). EDP Sciences.

BELIAO J., *(2014), Characterizing Speech Genres through the Relation between Prosody and Macrosyntax.* In Pristine Perspectives on Logic, Language, and Computation (pp. 1-18). Springer Berlin Heidelberg.

BELIAO J., LACHERET A., (2013) *Disfluencies and discursive markers : when prosody and syntax plan discourse*, The 6th Workshop on Disfluency in Spontaneous Speech (DISS2013), KTH Royal Institute of Technology, Stockholm, Suède.

BERRENDONNER A. (1990) , Pour une macro-syntaxe. Travaux de linguistiques, 21, 25-36.

BIBER D., JOHANSSON S., LEECH G., CONRAD S., FINEGAN E. (2000), *Grammar of Spoken and Written English*, Harlow, Longman.

BIBER D., CONRAD S. (2009), Register, genre, and style. Cambridge University Press.

BENZITOUN C., DISTER A., GERDES K., KAHANE S., PIETRANDREA P., SABIO F. (2010), t*u veux couper là faut dire pourquoi : propositions pour une segmentation syntaxique du français parlé, Actes du 2ème congrès mondial de linguistique française*, CMLF2012, La Nouvelle Orléans, 2075-2090.

BLANCHE-BENVENISTE C., JEANJEAN C. (1986), *Le français parlé. Editions et transcription.* Paris: Didier Erudition.

BLANCHE-BENVENISTE C., BILGER M., ROUGET Ch. & VAN DEN EYNDE K. (1990), Le français parlé : études grammaticales, Paris, Editions du CNRS.

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE Fl., PIRES M., (2012), *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000* (CFPP2000) http://cfpp2000.univ-paris3.fr/.

BREIMAN, L.,(1984), *Classification and regression trees*, Chapman & Hall/CRC.

CRESTI, E., (2000) Corpus du italiano parlato : Introduzione, Academia della Crusca, 1.

DETEY, S., DURAND J., LAKS B. & LYCHE C. (éds.), (2010) *Les variétés du français parlé dans l'espace francophone. Ressources pour l'enseignement*. Paris: Ophrys.

DURAND J., LAKS B. & LYCHE C. (2009), *Le projet PFC (phonologie du français contemporain): une source de données primaires structurées*, in : J. Durand, B. Laks & C. Lyche (eds.), *Phonologie, variation et accents du français*. Hermès, Paris, 19-61.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C., TELLIER I., (2012), Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012., in Ressources linguistiques libres, TAL. Volume 52 – n° 3/2011, 17-46.

GOLDMAN J.-Ph. (2011). Easyalign: an automatic phonetic alignment tool under praat. *Proceedings INTERSPEECH*, 3233-3236. http://latlntic.unige.ch/phonetique.

JAKOBSON, R. (1963), *Linguistique et poétique*, Essais de linguistique générale, Paris, Minuit, p. 209-248.

JOLLIFFE I.T., (2002) *Principal Component Analysis,* Springer Series in Statistics.

HABERT B., NAZARENKO A., SALEM A. (1997). *Les linguistiques de corpus*, Paris, Armand Colin.

HALLIDAY M. (1989), *Spoken and Written Language*, second edition, Oxford, Oxford University Press.

HOTELLING H.,(1933), *Analysis of a Complex of Statistical Variables with Principal Components*, Journal of Educational Psychology, Vol. 24 , pp. 498-520.

KAHANE S., PIETRANDREA P. (2009), *Les parenthétiques comme « Unités Illocutoires Associées » : une perspectove macrosyntaxique*, in M. Avanzi & J. Glikman (éd.), *Les Verbes Parenthétiques : Hypotaxe, Parataxe ou Parenthèse ?*, Linx, 61, 49-70. (paru en 2012).

KOCH, P. and W. OESTERREICHER (2001). *Langage parlé et langage écrit*, in Lexicon der Romanistischen Linguistik, T1-2, Tübingen, Max Niemeyer Verlag, 584-627.

LACHERET-DUJOUR A., VICTORRI B. (2002), *La période intonative comme unité d'analyse pour l'étude du*

français parlé : modélisation prosodique et enjeux linguistiques, Verbum n°1-2 : Y-a-t-il une syntaxe au-delà de la phrase ? M. Charolles, P. Le Goffic & M.A. Morel (Eds.), Nancy, 55-72.

LACHERET A., OBIN N., AVANZI M. (2010), *Design and evaluation of shared prosodic annotation for spontaneous French speech: from expert knowledge to non-expert annotation* , Proceedings 48th Annual Meeting of the Association for Computational Linguistics, 4th Linguistic Annotation Workshop, juillet 2010, 265-273.

LACHERET A., KAHANE S. PIETRANDREA P, AVANZI M., VICTORRI B. (2011), *Oui mais elle est où la coupure là ? Quand syntaxe et prosodie s'entraident ou se complètent* , Langue française, 170 : Unités syntaxiques et unités prosodiques, Fl. Lefeuvre & E. Moline (Eds.), Paris-Larousse, 61-80.

LACHERET-DUJOUR A. KAHANE S., PIETRANDREA P. (éds.), (à par. 2015a), R*hapsodie: a Prosodic and Syntactic Treebank for Spoken French*, Amsterdam, Benjamins.

LACHERET-DUJOUR Anne, BORDAL Guri, TRUONG Arthur (à par. 2015b), *The prosodic Structure*, Lacheret *et al.*

MAINGUENEAU D. (1996), L*es termes clés de l'analyse du discours*, Paris, Seuil.

MERTENS P. (1987) , *L'intonation du français : de la description linguistique à la reconnaissance automatique,* Thèse de Doctorat, Université de Louvain.

MOREL M.-A., DANON-BOILEAU L., (1998), *Grammaire de l'intonation*, Gap-Paris, Ophrys.

PEARSON K., (1901–36) *Biometrika*, Vol. 100, No. 1., pp. 3-15.

SWALES J.M. (1990), *Genre Analysis,* Cambridge: Cambridge University Press.

TERKEN, John. (1991), *Fundamental Frequency and Perceived Prominence*, Journal of the Acoustical Society of America, 89, 1768-1776.

WITHERS P. (2012), *Metadata Management with Arbil*, In Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme (p. 72).