Identification de descripteurs pour la caractérisation de registres

Jade Mekki, Delphine Battistelli, Gwénolé Lecorvé, Nicolas Béchet

Identification de descripteurs pour la caractérisation de registres

Jade Mekki[1,2] Delphine Battistelli[2] Gwénolé Lecorvé[1] Nicolas Béchet[3]
(1) Univ Rennes, CNRS, IRISA, 6, rue de Kerampont, 22305 Lannion Cedex, France
(2) Universite Paris-Ouest-Nanterre, MoDyCo, 200, avenue de la République 92001 Nanterre Cedex, France
(3) Universite de Bretagne Sud, IRISA, Campus de Tohannic, rue Yves Mainguy, 56017 Vannes Cedex, France
prenom.nom@irisa.fr, delphine.battistelli@u-paris10.fr

RÉSUMÉ

L'article présente une étude des descripteurs linguistiques pour la caractérisation d'un texte selon son registre de langue (familier, courant, soutenu). Cette étude a pour but de poser un premier jalon pour des tâches futures sur le sujet (classification, extraction de motifs discriminants). À partir d'un état de l'art mené sur la notion de registre dans la littérature linguistique et sociolinguistique, nous avons identifié une liste de 72 descripteurs pertinents. Dans cet article, nous présentons les 30 premiers que nous avons pu valider sur un corpus de textes français de registres distincts.

ABSTRACT

Feature identification for register characterization.

The paper presents a study of linguistic features for the characterization of a text according to its language register (formal, neutral, informal). This study aims at laying a first milestones for future work on this subject (e.g., classification, discriminating patterns extraction, etc.). From a state of the art conducted on the notion of register in linguistics and sociolinguistics, we have identified a list of 72 relevant descriptors. In this paper, we present the first 30 ones that we could validate on a corpus of French texts from distinct registers.

MOTS-CLÉS : registres de langue, descripteur linguistique, validation.

KEYWORDS: language register, linguistic feature, validation.

1 Introduction

Cet article présente les premiers travaux liés à un projet plus vaste dont la finalité est d'identifier automatiquement le registre d'un texte, puis de générer des paraphrases qui le transposent vers un registre différent. La caractérisation automatique d'un registre n'a, à notre connaissance, pas fait l'objet de travaux en TAL. Nous situant pour notre part dans cette perspective, une première brique de notre travail consiste à pouvoir disposer d'une première liste de descripteurs suffisamment exhaustive et validée en corpus. Nous exposons donc un travail de recherche exploratoire qui tend à identifier des descripteurs linguistiques pour différents registres (familier, courant et soutenu). C'est cette première brique que nous décrivons ici. L'analyse s'appuie conjointement sur un travail d'expertise et des résultats statistiques. Cette phase préparatoire nous a permis d'associer certains motifs à un registre particulier : nous tendons simplement à exposer les descripteurs étudiés ainsi que leurs comportements en contexte.

Nous commençons ici par revenir sur la notion même de registre en linguistique (section 2), avant de

présenter notre méthodologie d'analyse et de validation ainsi que notre corpus (section 3). Quelques descripteurs sont détaillés (section 4) avant de donner les résultats en (section 5), accompagnés d'une discussion.

2 La notion de registre de langue

Chaque production linguistique est évaluée par l'interlocuteur. Il la caractérise en la situant dans une classe, un registre. Ce dernier permet de qualifier une certaine actualisation de la langue. Ainsi, la notion de registre de langue se situe à la jonction de la linguistique et de la sociolinguistique. Elle s'entend globalement comme renvoyant la variété linguistique associée à une situation de communication particulière, indépendamment de paramètres liés au locuteur/scripteur comme, par exemple, son origine sociale ou son état émotionnel, et se caractérise par des patrons linguistiques spécifiques (Ferguson, 1982; Ledegen & Léglise, 2013). Néanmoins, les travaux du domaine mettent en exergue une difficulté définitoire. Ainsi, les dénominations « niveau », « style » ou encore « genre » co-existent avec celle de « registre » et font l'objet de débats (Bell, 1984; Biber & Finegan, 1994; Gadet, 1996). Ainsi, le partitionnement de l'espace linguistique en registres va varier selon l'angle d'étude. Par exemple, Ilmola (2012) mettra l'accent sur les registres familier, populaire et vulgaire dans des journaux satiriques, là où Borzeix & Fraenkel (2005) s'intéresseront à catégoriser différentes situations de communication au travail pour mettre notamment en exergue la manière dont les mots permettent de mettre en place une forme de contestation des normes établies. De fait, quand on aborde la notion de « registre », la question de (l'écart à) la norme apparaît comme centrale.

En effet, la perception d'un registre passe par l'évaluation de la façon de parler d'un locuteur. Bien que le sujet de notre étude ne soit pas de discuter des ressorts qui associent telle production avec telle connotation, il existe de fait une évaluation axiologique de chaque production linguistique : une forme de jugement de valeur assortie d'une hiérarchisation inhérente à toute évaluation de production linguistique.

Cette notion de hiérarchie nous amène à interroger celle de la « norme » : comment pouvons nous la définir ? Le registre « courant » semble représenter ce que nous appelons la norme. Effectivement, instinctivement nous pourrions penser qu'elle est constituée de toutes les productions linguistiques qui suivent correctement les règles grammaticales françaises. Cependant, une seconde interprétation de la norme serait qu'elle ne produise pas l' évaluation mais qu'au contraire, elle la renforce grâce à son mécanisme de rationalisation. Autrement dit, elle donnerait des justifications a posteriori de ce que nous avons perçu comme étant une variation. La seconde proposition permet de dépasser la valeurs normative a priori d'un bon usage de la langue. Cependant, cette définition est difficilement formalisable en traitement automatique des langues : c'est pourquoi nous adoptons la première proposition.

Par ailleurs, il est souvent admis que le partitionnement en registres distincts est davantage une commodité théorique qu'une réalité de terrain, les différentes pratiques de la langue s'exprimant généralement selon un continuum et non des oppositions tranchées (Blanche-Benveniste, 1997). Notre travail ne visant pas – en tout cas de manière directe – à une contribution sur ces questions, nous adoptons volontairement une approche consensuelle en employant le terme « registre », issu de la tradition britannique (Ure, 1982; Sanders, 1993), et en distinguant trois grands registres communément admis : familier, courant et soutenu. Si nous constatons et admettons qu'un continuum existe, nous devons en effet envisager la notion de registre avec des valeurs discrètes afin d'appréhender cette

notion d'un point de vue automatique.

Un examen approfondi de la littérature linguistique et sociolinguistique nous a permis de répertorier un nombre relativement important de caractéristiques linguistiques, de nature différente (lexicale, syntaxique, etc.), classiquement retenues et sur lesquelles nous reviendrons plus en détail dans la section 5. Les exemples (1), (2), (3), (4) donnent un aperçu de la complexité plus ou moins grande qu'il peut y avoir à envisager pour classifier automatiquement des phrases ou des textes dans un registre.

(1) *Moi, les enquêtes de terrain, bof.* tiré de (Frei, 1929)

(2) *Aussi dément que cela paraisse, il prend au sérieux le droit à l'alimentation, sur une terre où 1,2 milliard [sic] de couillons loin de chez nous souffrent d'une faim chronique.* tiré de (Ilmola, 2012).

(3) *Vous détenez un petit compte bancaire aux Bahamas ou à Jersey ? Comment puiser discrètement, et à distance, dans ce magot sans être tracassé ?* tiré de (Ilmola, 2012).

(4) *Le conseil général la saque non pas parce qu'elle gagne un peu trop, mais au contraire pour ne pas avoir atteint un « revenu minimum » d'au moins 701 euros par mois avec son activité d'autoentrepreneur !* tiré de (Ilmola, 2012).

Dans ces quatre exemples seul un terme par phrase est du registre familier voire vulgaire : « bof », « magot », « saque », « couillons ». Il est intéressant de noter que le registre neutre oscille avec le soutenu car nous trouvons la forme « cela » dans le troisième exemple, ainsi que des planificateurs de discours dans le deuxième : « non pas parce qu'... », « mais au contraire pour… ». En outre, il est important de mettre en regard le but de ces citations et leur lexique : ce sont des extraits tirés de journaux satiriques qui tendent à faire rire leurs lecteurs. Or l'effet comique naît d'une rupture de l'horizon d'attente du lecteur (Bergson, 2013). Dès lors, le décalage entre le genre neutre voire soutenu de la phrase avec un élément lexical du registre familier voire vulgaire tend à opérer cette rupture afin de créer l'effet comique. Ce mécanisme ne met pas réellement en lumière une variation linguistique mais au contraire une utilisation consciente des registres afin de jouer avec la perception de ces derniers chez le lecteur dont la surprise vient de la capacité à identifier les différents registres mis en oeuvre. Ainsi, nous trouvons dans ces exemples un style mimétique de différents registres sciemment utilisés : dès lors, un registre mimé reste il valable ? Ou bien est ce l'illustration de notre perception de ce registre qui est exposé ?

Nous ne visons toutefois pas à répondre à ces questions car notre objectif à terme est de pouvoir évaluer et classer automatiquement une production linguistique, en l'occurrence un texte entier et non un terme ou une phrase isolée seulement. Considérant les registres familier, courant et soutenu, notre but est de construire un jeu de descripteurs (terme que nous définirons plus bas) susceptibles d'être utiles pour cela, c'est-à-dire un ensemble d'éléments quantifiés portant sur des propriétés linguistiques.

3 Présentation de la méthodologie

Nous avons choisi de partir de caractéristiques linguistiques identifiées dans la littérature, que nous avons catégorisées, puis complétées. Nous avons ensuite cherché à valider ou invalider en corpus la nature réellement discriminante de chacune d'entre elles pour les registres considérés. Cette section

présente la méthode de validation, puis le corpus que nous avons construit.

3.1 Validation d'un descripteur

Notre approche de validation s'appuie sur des comparaisons de fréquences d'une caractéristique linguistique entre corpus associés à chaque registre. Pour un corpus donné, chaque caractéristique étudiée (par exemple, l'emploi du mot « ça ») est décrit par sa fréquence d'apparition relative dans un texte, c'est-à-dire normalisée par la longueur en mots du corpus. Par abus de langage, nous abrègerons nos propos en parlant simplement de « descripteur X » pour faire référence à la « fréquence relative d'apparition de la caractéristique X ». Considérant trois corpus textuels, chacun spécifique à un registre, nous posons alors un descripteur comme valide pour un registre donné si, parmi les différents corpus, la valeur du descripteur est significativement [1] supérieure pour le corpus dédié à ce registre à celle des autres.

Cette approche est volontairement simpliste pour rester indépendante d'un maximum d'hypothèses. Notre travail ne prétend ainsi pas statuer de manière absolue sur la validité de tel ou tel descripteur mais dresse un panorama du niveau de fiabilité d'un large panel de descripteurs. Cet étalonnage a pour finalité d'offrir un point de départ et de comparaison à de futurs travaux plus avancés, par exemple des techniques d'analyse plus fines et des motifs extraits automatiquement par des méthodes de fouille. L'approche par fréquence donne en effet différents niveaux de lectures intéressants. D'une part, outre la validité pour un registre, elle permet également de donner la non validité pour un registre dans le cas où un descripteur est significativement inférieur pour ce registre que pour les deux autres. D'autre part, l'analyse contrastive peut être affinée en partitionnant le corpus de chaque registre (Efron, 1979). Ainsi, le nombre de partitions pour lesquelles un descripteur est significativement supérieur permet de quantifier la fiabilité d'un descripteur. Dans cet article, nous présentons des résultats polarisés mais ne reportons pas de niveaux de confiance.

3.2 Corpus

3.2.1 Constitution et traitement du corpus

Les trois types de corpus considérés sont composés d'écrits français : *Albertine disparue* de Proust pour le registre soutenu, des archives de *L'Humanité* et *Le Monde* pour le registre courant et *Kiffe Kiffe Demain* de Guene, *L'Assommoir* de Zola et *Voyage au bout de la nuit* de Céline pour le registre familier. Les trois types de corpus contiennent respectivement $110\,000$, $2\,500\,000$ et $370\,000$ mots. Ces données ont été tokenisées, puis étiquetées en parties du discours avec Treetagger.

3.2.2 Questions soulevées par le corpus

Comme précédemment évoquée, la norme (section 2) peut être envisagée comme le respect des règles grammaticales. Cette première interprétation met en exergue un paradigme où la norme est associée à l'écrit. Par exemple, pour apprendre à maîtriser le français nous faisons des dictées non des improvisations orales : la standardisation interviendrait alors à l'écrit tandis que la variation se

1. S'agissant d'un travail préliminaire, ce critère de significativité n'a pas encore été formalisé.

produirait à l'oral. Ce paradigme sous-tend toute la notion de « registre » puisque l'évaluation opère selon la norme. Ainsi, plus la production semble orale plus elle va être perçue comme familière car spontanée et non structurée, en revanche plus elle paraît écrite plus elle va être appréhendée comme soutenue car perçue plus construite et réfléchie (donc plus rationnelle).

Dès lors, notre corpus est problématique puisque son medium est l'écrit quelque soit son registre. Autrement dit, la parfaite maîtrise des registres par Zola ou Céline, par exemple, invalide-t-il le fait que *L'Assommoir* ou bien *Voyage au bout de la nuit* soient considérés comme du registre familier ?

Toutefois, les travaux présentés se situent au début du projet et ne tendent pas à répondre à ces questions mais à les soulever afin d'explorer plusieurs axes de réflexion.

4 Descripteurs répertoriés

Nous avons dressé une liste de 72 descripteurs, émanant soit de la littérature (66), soit d'une analyse préliminaire conduite par nos soins pour les différents registres (6 descripteurs supplémentaires ont ainsi été identifiés pour le registre familier). Nous sommes conscients de la non exhaustivité de cette liste qui sera amener à évoluer dans de futurs travaux. Ces descripteurs sont généralement soit des motifs « fautifs » (au sens d'un écart à la norme), soit des motifs corrects (toujours selon la norme) mais rares. Ces descripteurs couvrent divers niveaux d'abstraction de la langue que nous regroupons sous les catégories lexicale (16 descripteurs), morphologique (16), syntaxique (38) et phonétique (2).

Dans l'absolu, certains indices lexicaux sont évidemment très discriminants (appartenance explicite de certains lexèmes à un registre donné). Nous n'avons cependant pas traité cet aspect malgré l'important potentiel discriminant de ce type de descripteur : il faudrait disposer de dictionnaires suffisamment exhaustifs [2]. De plus, nous n'avons pas eu recours à une mesure de richesse lexicale car cette notion nous a semblé délicate à traiter. De fait, plus il y a de différents termes lexicaux employés, plus nous pouvons supposer que le vocabulaire est riche donc soutenu. Toutefois, le registre familier est également reconnu pour sa créativité. Pour être efficace, la mesure lexicale devrait ainsi s'appuyer sur une distinction entre termes standards et exotiques, par exemple sur la base d'un dictionnaire à nouveau. Enfin, notons que l'étude de descripteurs phonétiques fait sens y compris pour une analyse de textes écrits (et non transcrits) car l'usage écrit de formes orales est désormais répandu à travers des modes de communications connectés (chats, messageries, textos…).

Quelques exemples sont donnés dans les sous-sections suivantes.

4.1 Phonétique

4.1.1 Elision du « e »

Voyage au bout de la nuit

1. J' vais te dire...
2. J' veux bien payer...
3. J' sais pas encore...

2. Des dictionnaires tels que wiktionnaire par exemple renseignent sur les registres de mots.

4. J’ m’en fous, j’irai me donner.

4.2 Lexical

4.2.1 Sur présence de « là » - ponctuant : familier

Les trois textes associés au registre familier (*L’Assommoir*, *Kiffe Kiffe Demain*, *Voyage au bout de la nuit*) sont ceux où le motif est le plus présent. Suite à une première recherche nous avons voulu préciser la place du motif dans la phrase : un facteur de sa dimension discriminante est sa fonction de « ponctuant ». De fait, lorsque nous avons cherché le motif lorsqu’il était en fin de phrase seuls deux des trois textes catégorisés comme « familiers » (*L’Assommoir*, *Kiffe Kiffe Demain*) ont une fréquence relative plus haute que les autres registres. Il est intéressant de noter que le troisième texte où le motif est le plus présent est du registre soutenu. Nous pourrions, dans des travaux futurs, affiner notre test et rechercher le motif uniquement dans des passages de discours rapporté par exemple.

4.3 Morphosyntaxique

4.3.1 Sujet « nous » transposé en « on »

La présence du sujet « on » est nettement plus présent que « nous » dans les trois oeuvres dont le registre est familier. A l’inverse, c’est dans *Albertine Disparue* associé au registre soutenu que « nous » est le plus présent.

Afin de mieux comprendre le comportement du motif, nous l’avons observé en contexte dans deux oeuvres associées au registre familier.

L’Assommoir :

1. On :
   (a) On la rencontrerait une nuit sur un trottoir, pour sûr.
   (b) Alors, comme on ne parlait pas toujours de leur mariage, elle voulut s’en aller, elle tira légèrement la veste de coupeau.
2. Nous :
   (a) … j’ai quelque chose à laver, je vous garderai une place à côté de moi, et nous causerons.
   (b) — Ca suffit entre nous, madame Gervaise, murmura-t-il.

Kiffe Kiffe Demain :

1. On :
   (a) On lui crie après sans arrêt, et on la surveille pour vérifier qu’elle pique rien dans les chambres.
   (b) Cette meuf, on dirait qu’elle a besoin d’être heureuse à la place des autres.
2. Nous :
   (a) Une fois, il a dit à ma mère qu’en dix ans de métier, c’était la première fois qu’il voyait " des gens comme nous avec un enfant seulement par famille "

(b) Ma mère, elle dit que si mon père nous a abandonnées, c'est parce que c'était écrit.

Ainsi, « on » permet au locuteur de rester impersonnel, ce qui est redoublé par son association à des termes qui généralisent le propos :

1. « on » + « pour sûr »
2. « on » + « toujours »
3. « on » + « sans arrêt »
4. « on » + maxime au présent de vérité général « elle a besoin d'être heureuse à la place des autres »

Tandis que « nous » a tendance à identifier des locuteurs précis :

1. « moi » + vous = « nous »
2. « madame Gervaise » + locuteur = « nous »
3. « ma mère » + « ma » (1 personne du SG) = « nous »
4. « ma mère » + « ma » / « mon » (1 personne du SG = « nous »

4.3.2 Syntagme : « ça + VB »

Lorsque nous observons le comportement du syntagme en contexte, nous constatons que dans les oeuvres associées au registre soutenu et courant, les motifs se trouvent dans des situations de prise de parole explicite avec des marqueurs d'oralité tels que des verbes de parole, les guillemets, la première personne du singulier...

Albertine disparue

1. « (...) si ça amuse le pauvre Swann de faire des bêtises et de ruiner son existence, c'est son affaire, mais on ne se prend pas avec ces choses-là, tout ça peut très mal finir (...) »
2. Et elle ajouta : « Ca devait arriver (...) »

L'Humanité

1. Il pense à « comment ça doit fonctionner ».
2. nous disait : « (...) Et ça va vite, trop vite pour nous.(...) »

Le Monde

1. « D'ordinaire, un conseiller ministériel, petite main e l'ombre, ça ferme sa gueule. » Pierre Jacquemain explique aux « Monde » son départ.
2. le président de la république a défendu ses réformes économiques et martelé que « ça va effectivement mieux pour la France. »

En revanche, dans les oeuvres associées au registre familier, le motif se trouve dans des situations où il n'y a pas une prise de parole explicite.

L'Assommoir

1. Chez elle, ça entrait et ça sortait.
2. Vrai, ça faisait un fameux débarras.

Kiffe Kiffe Demain

1. Ca va rien couter à ta mère si c'est ça qui te préoccupe. De toute facçon, le ski ça pue la merde.
2. Ca marche bien.

Voyage au bout de la nuit

1. Quel effet que ça avait bien pu lui faire ?
2. Ca crève un homme...

Cela met en exergue le lien implicite entre le registre familier et l'oral lorsque ce dernier est représenté à l'écrit. Ce lien pose la question du style plutôt que du genre. En effet, le genre semble être motivé par un facteur extra-linguistique comme le besoin de désambiguiser (faire phonétiquement la différence entre le singulier et le pluriel « il croit » / « ils croivent ») par exemple. Or, ici le besoin semble justement d'être mimétique du genre. Autrement dit, le corpus refléterait l'utilisation de styles pour s'approcher d'un genre. La question de la validité des textes se pose à nouveau. Pour y répondre nous pourrions, dans de futurs travaux, poursuivre notre exploration en utilisant un nouveau corpus oral composé des différents registres.

5 Résultats et discussion

Nous présentons ici les 30 descripteurs ayant été d'ores et déjà validés dans notre corpus, les 42 restants nécessitant soit davantage de ressources textuelles (par extension du corpus initial), soit le recours à des outils plus ou moins complexes (ex. analyseur syntaxique) dont certains sont encore inexistants (par ex., un outil détectant les ellipses, cf. exemple (1), section 2).

La table 1 présente ces 30 descripteurs validés par niveau d'abstraction. Pour chacun, nous indiquons le registre pour lequel il a été validé comme positivement ($+$) ou négativement ($-$) discriminant, ainsi que la référence bibliographique d'où il est tiré. Ceux n'ayant pas de référence sont des descripteurs que nous avons nous mêmes proposés de prendre en considération.

Il ressort que la majorité des descripteurs validés concernent le registre familier. Ce constat s'explique par la créativité de ce registre vis à vis de la norme , qu'il s'agisse de la richesse lexicale puisant dans les lexiques populaires, argotiques voire vulgaires, ou de la multiplicité de manières de s'écarter de la norme (et donc de produire des "fautes"). La notion de faute est d'ailleurs intéressante, non pas pour sa valeur axiologique implicitement associée, mais pour les raisons qui amènent le locuteur à produire un énoncé fautif. Comme le souligne Frei (1929), « on ne fait pas des fautes pour le plaisir de faire des fautes. (...) Dans un grand nombre de cas la faute, qui est passée jusqu'à présent pour un phénomène quasi pathologique sert à prévenir ou à réparer les déficits du langage correct. » Autrement dit, la faute serait le symptôme d'un « déficit » du français. Il y aurait donc une sorte de régularisation spontanée des irrégularités arbitraires de la langue normée. Frei (1929) présente alors les fautes comme venant pallier un « besoin » : besoin de désambiguïser, besoin d'être expressif. . .

Registre	Source	F	C	S
Niveau lexical (5)				
Éléments ponctuants	(Gadet, 2003)	+		
Onomatopées	(Ilmola, 2012)	+		
« Là » ponctuants	(Gadet, 1997)	+		
Termes à redoublement (« tonton », « dodo »)	(Gadet, 1997)	+		
Planificateurs du discours (« néanmoins », « en raison de »)	(Branca-Rosoff, 1999; Bilger & Cappeau, 2004)			+
Niveau morphosyntaxique (13)				
Contraction de « cela » en « ça »	(Gadet, 1997)	+		−
Négation sans « ne »	(Bilger & Cappeau, 2004)	+		−
Sujet « on » transposé en « nous »	(Bilger & Cappeau, 2004)	−		+
Terminaison en « -asse »	(Ilmola, 2012)	+		
Terminaison en « -ouze »	(Ilmola, 2012)	+		
Terminaison en « -o »	(Ilmola, 2012)	+		
Verbe « être » au singulier devant un syntagme nominal singulier	(Bilger & Cappeau, 2004; Favart, 2011)	+		
« Ça » + verbe		+		
Dérivation en adverbe d'un nom ou adjectif (« vachement »…)	(Ilmola, 2012)	+		
Verbes du premier groupe	(Gadet, 1997)	+		
Emploi du passé simple		−		+
Emploi du passé composé			+	−
Emploi du présent de l'indicatif			+	
Niveau syntaxique (10)				
Emploi fautif de relatives en « que » (« relative populaire »)	(Gadet, 2003)	+		
Interrogative sans inversion sujet/verbe	(Gadet, 2003)	+		
Interrogative en « est-ce que »	(Ilmola, 2012)		+	
Maintien de « des » devant un adjectif au lieu de « de »	(Ilmola, 2012; Kalmbach, 2012)	+		
Rajout de « à lui/elle » après un pronom personnel « son/sa »	(Gadet, 2003)	+		
Emploi de pronoms relatifs (« dont », « lequel »…)	(Gadet, 2003)			+
Adverbe + parataxe (« vraiment bien »…)		+		
Inversion « en » et COI à l'impératif (« donne m'en »)	(Kalmbach, 2012)	+		
« C'est …qui » (« c'est lui qui a fait ça »)		+		
Effacement du pronom « il » impersonnel (« fallait pas… »)	(Favart, 2011)	+		
Niveau phonétique (2)				
Élision de « e » (Favart, 2011)		+		
Élision du « i » du pronom « qui » devant une voyelle	(Ilmola, 2012)	+		

TABLE 1 – Descripteurs validés positivement (+) et négativement (−) pour les registres familier (F), courant (C) et soutenu (S).

Il arrive que des fréquences relatives d'un même motif dans deux textes associés à des registres différents soient très proches. C'est le cas par exemple des « termes à redoublement ». On observe qu'ils ont la fréquence relative la plus importante dans Kiffe Kiffe Demain (associé au registre familier) mais que la seconde fréquence relative la plus haute se trouve dans Albertine Disparue (associé au registre soutenu). Dans ce cas précis, nous avons décidé malgré tout de valider ce motif en tant qu'évocateur du registre familier mais ceci a attiré notre attention et nous avons alors observé plus en détail le contexte du motif avec un concordancier. Nous avons notamment alors remarqué une corrélation entre la présence de discours rapportés (identifiés via la présence de verbes de parole ou de marques de ponctuation telles que les guillemets) et les descripteurs discriminants pour le registre familier, ceci semblant étayer le fait (qu'il faudrait bien entendu approfondir) qu'une certaine oralité serait plus étroitement associée au registre familier. Nous avons par ailleurs constaté que l'apparition d'un descripteur familier était souvent doublée par la présence d'autres descripteurs familiers : « c'est ça » (forme contractée + redondance sémantique : « cela est cela »), « ça rime à quoi ? » (forme contractée + non inversion de forme interrogative + phrase courte). Ainsi, un descripteur serait toujours renforcé par la présence d'un second descripteur. Nous pourrions donc, dans de futurs travaux, chercher à identifier des n-uplets de descripteurs plutôt que des descripteurs isolés.

D'un point de vue théorique, il est clair que nous nous plaçons dans une étude qui rend centrale – conjointement à celle de l'écart à la norme – la question du continuum oral-écrit, à l'instar des travaux de (Blanche-Benveniste, 1997), lesquels ont notamment souligné que ce qui est souvent présenté comme spécifique des modalités orale *vs.* écrite concerne (aussi) l'opposition entre registres formel *vs.* informel. Dans une méthodologie proche de celle de (Douglas, 1988), nos travaux devraient permettre de proposer un ensemble exhaustif - et nous l'espérons pertinent - de descripteurs linguistiques à même d'éclairer ces deux types d'oppositions.

6 Conclusion

Dans le travail présenté ici, nous sommes revenus sur des caractéristiques identifiées dans la littérature (au nombre de 72) abordant la notion de registre pour en proposer un mode de validation puis une catégorisation utile pour des tâches futures auxquelles nous allons nous atteler prochainement (classification, extraction de motifs discriminants). Le résultat de ce travail est ainsi une première liste de 30 descripteurs validés parmi 48 descripteurs testés.

Outre le fait de continuer l'investigation des 24 descripteurs restants non encore testés, un prolongement immédiat à ce travail concerne l'établissement de degrés de confiance associés aux descripteurs validés, par exemple par des techniques de *bagging* et d'analyse de la variance. Ce travail devrait également concerner les 18 descripteurs que nous avons testés mais que nos corpus n'ont pas permis de valider de manière certaine. De manière corrélée, un second axe de travail contribuant également à une vision plus fine des phénomènes étudiés est la constitution d'un corpus de grande ampleur et faisant intervenir de manière croisée et explicite tout à la fois les notions de genre (en nous appuyant sur les réflexions de Adam (1999)) et de registre (dans le sens où nous l'avons défini en section 2). Cette première liste de descripteurs validés oeuvre en ce sens car elle devrait permettre de construire un premier outil (règles expertes, classifieur simple) pour filtrer et annoter automatiquement de grands corpus (libres, journaux, pages web…).

Remerciements

Ce travail a bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR) dans le cadre du projet TREMoLo (ANR-16-CE23-0019).

Références

ADAM J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Nathan.

BELL A. (1984). Language style as audience design. *Language in society*, **13**(2), 145–204.

BERGSON H. (2013). *Le rire*. Flammarion.

BIBER D. & FINEGAN E. (1994). *Sociolinguistic perspectives on register*. Oxford University Press on Demand.

BILGER M. & CAPPEAU P. (2004). L'oral ou la multiplication des styles. *Langage et société*, (3).

BLANCHE-BENVENISTE C. (1997). Approches de la langue parlée en français. *Ophrys, Paris*.

BORZEIX A. & FRAENKEL B. (2005). *Langage et travail (communication, cognition, action)*. CNRS éd.

BRANCA-ROSOFF S. (1999). Des innovations et des fonctionnements de langue rapportés à des genres. *Langage et société*, **87**(1), 115–129.

DOUGLAS B. (1988). Variation across speech and writing. *Cambridge : CUP*.

FAVART F. (2011). Le stéréotype de registre de langue populaire dans le roman du second XXe siècle (1966-2006). In *Stéréotypes en langue et en discours*. Centre Interlangues.

FERGUSON C. A. (1982). Simplified registers and linguistic theory. *Exceptional language and linguistics*, p. 49–66.

FREI H. (1929). *La grammaire des fautes : introduction à la linguistique fonctionnelle, assimilation et différenciation, brièveté et invariabilité, expressivité*, volume 1. Slatkine.

GADET F. (1996). Variabilité, variation, variété : le français d'europe. *Journal of French Language Studies*, **6**(1), 75–98.

GADET F. (1997). La variation, plus qu'une écume. *Langue française*, p. 5–18.

GADET F. (2003). Is there a french theory of variation ? *International Journal of the Sociology of Language*, **165**.

ILMOLA M. (2012). Les registres familier, populaire et vulgaire dans le canard enchaîné et charlie hebdo : étude comparative.

KALMBACH J.-M. (2012). *La grammaire du français langue étrangère pour étudiants finnophones*. Kielten laitos, Jyväskylän yliopisto.

LEDEGEN G. & LÉGLISE I. (2013). *Variations et changements linguistiques*. ENS Editions.

SANDERS C. (1993). *Sociosituational variation*. Cambridge : Cambridge University Press.

URE J. (1982). Introduction : approaches to the study of register range. *International Journal of the Sociology of Language*, **1982**(35), 5–24.