Descripteurs Linguistiques et Caractérisation Objective des Catégories Textuelles

Marina Seghier, Alice Millour, Jean-Yves Antoine

Descripteurs Linguistiques et Caractérisation Objective des Catégories Textuelles

Marina Seghier[1] Alice Millour[1] Jean-Yves Antoine[2 3]
(1) LIASD - Université Paris 8, 2 rue de l'Université, 93526 Saint-Denis, France
(2) LIFAT - Université de Tours, 64 avenue Jean Portalis, 37200 Tours, France
(3) LIFO - Université d'Orléans, 6 Rue Léonard de Vinci, 45067 Orléans, France
ms@up8.edu, am@up8.edu, jean-yves.antoine@univ-tours.fr

RÉSUMÉ

Nous présentons de premiers résultats s'inspirant des travaux de Biber (1988) utilisant des descripteurs linguistiques pour redéfinir les dimensions de la variation textuelle en français.

ABSTRACT

Linguistic Descriptors and Objective Characterization of Textual Categories

We present first results inspired by the work of Biber (1988) using linguistic descriptors to redefine the dimensions of textual variation in French.

MOTS-CLÉS : variation textuelle, classification non supervisée, descripteurs linguistiques, évaluation, annotation.

KEYWORDS: textual variation, unsupervised classification, linguistic descriptors, evaluation, annotation.

Les systèmes développés et de plus en plus répandus aujourd'hui, sont présentés comme étant très performant pour un grand nombre de tâches. Or, les performances annoncées ne sont pas toujours celles rencontrées selon les types (ou "genres") de ressources textuelles auxquels ces outils sont confrontés. En effet, on peut par exemple observer un différentiel de performances important entre différentes catégories de texte pour la tâche de reconnaissance d'entités nommées (Millour *et al.*, 2022) et d'étiquetage morpho-syntaxique. Cependant, les typologies textuelles existantes, fondées sur une classification des catégories *a priori* sans justification linguistique, ne permettent pas d'expliquer ce différentiel.

1 Précédents dans la Classification Textuelle

Biber (1988) a été le premier à mener une étude statistique sur ce qu'il appelle "genre" en anglais britannique, dans le but d'identifier plusieurs dimensions de la variation dans la langue. Il a procédé à une "analyse factorielle" (ACP) sur le LOB (*Lancaster-Oslo-Bergen*, un million de mots environ – composé de plusieurs échantillons de "genres" issus de l'écrit :

presse, religion, humour...), le *London-Lund* (500 000 mots environ – composé de textes de parole transcrite : entretiens, discours spontanés, préparés...), et une collection de ses propres lettres manuscrites.

Au terme de son analyse basée sur 67 caractéristiques calculées à partir de descripteurs linguistiques (tels que les pronoms personnels, les verbes au passé, la longueur des mots, des phrases, etc.), il a affirmé que la variation linguistique était continue selon six dimensions : 1) impliqué (affectivement) *VS* informationnel, 2) narratif *VS* non-narratif, 3) référence explicite *VS* dépendante de la situation, 4) expression manifeste de persuasion, 5) abstrait *VS* non-abstrait et 6) production d'informations sous contrainte temporelle.

À la manière de Passonneau *et al.* (2014), qui ont actualisé cette recherche pour l'anglais américain à partir du corpus MASC (*Manually Annotated Sub-Corpus*, 500 000 mots environ), nous proposons une caractérisation linguistique de différents genres à partir d'un corpus multisource pour le français.

2 Des Données à l'Analyse en Composantes Principales

Nous avons travaillé à partir de FENEC (*FrEnch Named-entity Evaluation Corpus*, 11 000 tokens environ – cf. table 1) (Millour *et al.*, 2022), un corpus d'évaluation pour la tâche de reconnaissance d'entités nommées en français. Il est composé de onze documents de six catégories textuelles différentes (poésie, prose, parole transcrite, encyclopédie, informations, multi-sources) et annoté manuellement en entités nommées selon divers jeux d'étiquettes.

Document	Période	Genre	Nb. phrases (Nb. tokens)	Licence
42131-0 (*Traité sur la Tolérance*, Voltaire)	XVIIIe	prose	40 (1 020)	Project Gutenberg
pg6470 (*Le Ventre de Paris*, Émile Zola)	XIXe		51 (1 002)	Project Gutenberg
pg6099 (*Les Fleurs du Mal*, Baudelaire)	XIXe	poésie	30 (1 014)	Project Gutenberg
56708-0 (*Œuvres d'Arthur Rimbaud - Vers et proses*)	XIXe		52 (1 027)	Project Gutenberg
UD French GSD	XXIe	multisources	35 (1 021)	CC BY-SA 4.0
Sequoia (Candito & Seddah, 2012)	XXIe		44 (1 002)	Licence LGPL-LR
French Question Bank (Seddah & Candito, 2016)	XXIe		102 (1 006)	Licence LGPL-LR
APIL (office du tourisme Othe-Armance)	XXIe	informations	29 (1 002)	Licence LGPL-LR
Wikinews	XXIe		46 (1 024)	CC BY 2.5
WikiNER français	XXIe	encyclopédie	36 (1 003)	CC BY 4.0
Spoken (Rhapsodie (Lacheret, Anne *et al.*, 2014))	XXIe	parole	70 (1 028)	CC BY-SA 4.0

TABLE 1 – Contenu du corpus annoté FENEC (échantillons de 1 000 tokens environ dans six genres) (Millour *et al.*, 2022).

Étant donné la taille peu conséquente de notre jeu de données, nous avons choisi de former des échantillons de 200 tokens, à partir desquels nous avons mesuré l'importance de certains descripteurs linguistiques pour le typage des textes, tels que les entités nommées, mais également les parties du discours (verbes, noms, déterminants, adjectifs...) et les verbes au

passé.

2.1 Choix et calcul des caractéristiques

Les caractéristiques utilisées pour cette expérience sont :

— les entités nommées du corpus FENEC (selon le schéma fin Quaero [1]), à savoir : PER, LOC, ORG, MISC (respectivement : personnes, lieux, organismes, divers) ;

— les parties du discours, pour lesquels nous avons testé plusieurs outils : 1) le modèle POET (*A French Extended Part-of-Speech Tagger*) de FLAIR [2], 2) CAMEMBERT [3] et 3) SPACY [4] ;

— les verbes au passé, annotés avec le *morphologizer* de SPACY [5].

Concernant l'étiquetage en parties du discours, après une brève comparaison des sorties produites par les trois outils, nous avons choisi de conserver les annotations du modèle POET de FLAIR pour calculer nos caractéristiques. Nous avons également réalisé une évaluation manuelle de ses résultats (environ 100 étiquettes par document) et de SPACY (toutes les occurrences de verbes annotés au passé, soit le rappel).

genre	Flair (précision)	SpaCy (rappel)
parole	78,43	58,14
encyclopedie	94,64	86,67
informations	90,29	77,42
prose	88,56	77,12
poesie	90,14	37,29

TABLE 2 – Résultats de l'évaluation manuelle de FLAIR et SPACY.

Sur la base de 26 caractéristiques calculées, nous avons choisi de mener notre expérience sur 23 d'entre elles, en retirant :

— MISC, soit les entités nommées "diverses" (*miscellaneous* en anglais) car ces dernières étaient trop variées au sein de cette catégorie (par exemple "boucher", "matelots", "enchanteresse", "*New York Times*", "*Hubble*", "*Surface and Depth*", "bataille d'Actium", "accords sur le charbon et l'acier"), et nous n'aurions pas été en mesure d'interpréter précisément et de manière fiable, l'impact de cette caractéristique sur l'ACP ;

1. Guide d'annotation : http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf
2. https://huggingface.co/qanastek/pos-french-camembert-flair
3. https://huggingface.co/gilf/french-camembert-postag-model
4. https://spacy.io/models/fr#fr_core_news_sm
5. https://spacy.io/models/fr

— X, soit les mots inconnus, tels que des mots anglais, des consonnes euphoniques ("t" dans "appelle-t-on"), "j" dans "j∼ j∼ j∼" (répétition dans la parole transcrite), car nous craignions un effet fort d'échantillonage, du fait du nombre total d'observables de cette étiquette dans le corpus, par rapport à la taille de celui-ci ;
— SYM, soit les symboles tels que le tilde, le pourcent, les symboles monétaires, l'abréviation de "environ" (env), pour la même raison que précédemment.

Dans la table 3, nous pouvons voir les occurrences d'un échantillon des observables pour 1 000 tokens.

	prose	parole	informations	encyclopedie	poesie
LOC	4	14	35	41	6
ORG	1	1	13	6	0
PER	23	5	8	22	11
TOTAL_EN	39	28	74	86	25
ADP	112	114	160	151	120
SCONJ	22	18	3	3	12
CCONJ	21	42	25	26	40
ADV	67	118	31	30	49
PROPN	23	27	66	93	17
NUM	10	19	54	57	3
AUX	37	42	22	35	19
VERB	125	135	75	66	82
DET	125	104	142	139	162
ADJ	110	118	123	137	156
NOUN	155	154	226	193	219
PRON	93	100	24	19	52
PPER1S	5	20	0	0	8
PPER2S	1	5	0	0	4
PPER3	25	16	8	7	7
INTJ	1	7	0	0	1
PUNCT	104	90	94	90	123
PAST_TENSE	59	43	36	45	30

TABLE 3 – Nombre d'occurrences d'un échantillon d'étiquettes pour 1 000 tokens, par genre.

2.2 Dimensions en français

D'après cette première analyse, une dimension semble se dessiner. D'une part, les textes de la catégorie 'poésie' sont davantage caractérisés par des mots plus longs et des noms détaillés, et précisés par des adjectifs ; ceux des catégories 'prose' et 'parole', par des conjonctions de subordination, des verbes en général, des verbes au passé, et à la voix passive. La présence d'entités nommées semble quant à elle caractériser davantage les genres 'informations' et 'encyclopédie'. En effet, nous pouvons supposer que les textes d'informations et d'encyclopédies présentent plus d'éléments tels que des zones géographiques, lieux d'intérêt, dates historiques, numéros de téléphones, etc.

Notre première composante dans la figure 1 présente des similitudes avec les cinquième

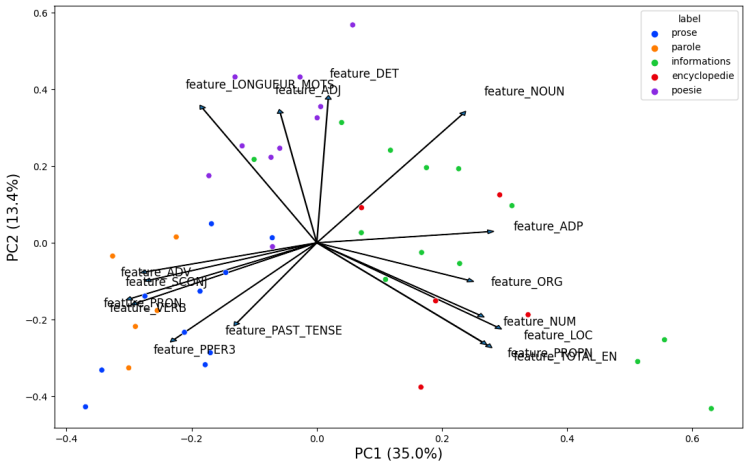

FIGURE 1 – Première et deuxième composantes principales, 23 caractéristiques.

de Biber (1988) et quatrième de Passonneau *et al.* (2014) (abstrait VS non-abstrait). Les catégories de productions écrites, détaillées d'un côté (poésie, prose) semblent plus conceptuelles (abstraites) ; et des catégories explicatives de l'autre (informations, encyclopédie) plus concrètes (non-abstraites).

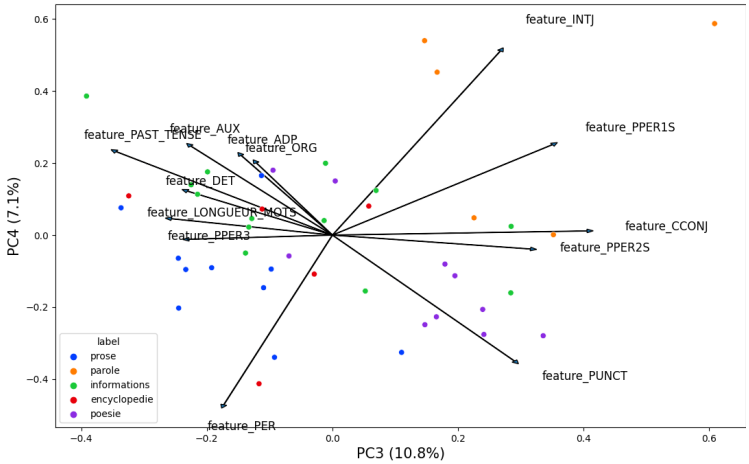

FIGURE 2 – Troisième et quatrième composantes principales, 23 caractéristiques.

D'après cette deuxième analyse, une dimension semble à nouveau se dessiner. D'une part, les textes de parole transcrite semblent caractérisés par des interactions à la 1ère et 2ème personne du singulier, et la présence d'interjections et de conjonctions de coordination. D'autre part, les textes en prose présentent davantage d'entités nommées de personnes, de pronoms à la 3ème personne et de ponctuation. Dans les textes d'informations, nous pouvons

trouver de la narration d'évènements passés, des mots plus longs et davantage de noms d'organismes.

Notre troisième composante dans la figure 2 présente ainsi des similitudes avec les deux premières de Biber (1988) et Passonneau *et al.* (2014) (impliqué VS non-impliqué). En effet, nous retrouvons d'un côté des textes d'interactions orales (parole) plus impliquées, et de l'autre des textes plus informationnels (encyclopédie, prose).

Contrairement aux deux premières composantes, celles-ci nous permettent de discriminer la parole de la prose, mais dans le cadre de cette étude, elles ne sont tout de même pas suffisantes pour discriminer plus nettement chaque catégorie textuelle, notamment celle de l'encyclopédie.

3 Conclusion

Bien qu'il reste une notion de continuum à approfondir, cette étude a permis de retrouver certaines similitudes avec les dimensions textuelles que Biber (1988) et Passonneau *et al.* (2014) avaient fait émerger en anglais britannique et américain, mais pour la première fois, sur un jeu de données multi-catégories français.

En outre, nous avons pu également éprouver trois outils : SpaCy pour à la fois l'annotation en parties du discours et en traits morphosyntaxiques, CamemBERT et le modèle POET de Flair pour l'annotation en parties du discours. L'évaluation de ces derniers nous a permis de constater sur ces deux tâches, ce qu'avaient observé Millour *et al.* (2022) sur la tâche de REN pour le français : un différentiel de performances important entre les catégories textuelles, avec des faiblesses majeures sur des textes de parole transcrite et de poésie.

Le cadre expérimental mis en place dans notre étude est modulable et permet facilement d'enrichir l'expérience avec d'autres caractéristiques, d'autres types de textes, et de donner lieu à de nombreuses extensions. L'une d'elles consiste à s'inspirer des travaux de Fu *et al.* (2020) et d'approfondir la notion des caractéristiques des entités nommées. Car au-delà de leurs types, ce sont sûrement leurs propriétés intrinsèques (longueur en caractères et en tokens, fréquence, ambiguïté, persistance...) qui mettent en difficulté les outils.

Références

Biber D. (1988). *Variation across Speech and Writing*. Cambridge University Press. DOI : 10.1017/CBO9780511621024.

Candito M. & Seddah D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In *Proceedings of*

the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, p. 321–334, Grenoble, France : ATALA/AFCP.

FU J., LIU P. & NEUBIG G. (2020). Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6058–6069, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.489.

LACHERET, ANNE, KAHANE, SYLVAIN, BELIAO, JULIE, DISTER, ANNE, GERDES, KIM, GOLDMAN, JEAN-PHILIPPE, OBIN, NICOLAS, PIETRANDREA, PAOLA & TCHOBANOV, ATANAS (2014). Rhapsodie : un treebank annoté pour l´étude de l´interface syntaxe-prosodie en français parlé. *SHS Web of Conferences*, **8**, 2675–2689. DOI : 10.1051/shsconf/20140801305.

MILLOUR A., DUPONT Y., JOUGLAR A. & FORT K. (2022). FENEC : un corpus équilibré pour l'évaluation des entités nommées en français (FENEC : a balanced sample corpus for French named entity recognition). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 82–94, Avignon, France : ATALA.

PASSONNEAU R. J., IDE N., SU S. & STUART J. (2014). Biber redux : Reconsidering dimensions of variation in American English. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 565–576, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

SEDDAH D. & CANDITO M. (2016). Hard time parsing questions : Building a QuestionBank for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2366–2370, Portorož, Slovenia : European Language Resources Association (ELRA).