Projet M2 TAL-IL (2020 – 2021)
Analyses et évaluations de motifs séquentiels émergents

Nicolas Béchet - nicolas.bechet@irisa.fr[1] and
Jade Mekki - jade.mekki@irisa.fr[1,2]

[1]IRISA
[2]CNRS-UPN MoDyCo

November 2020

Abstract

Dans le cadre du cours d'apprentissage automatique du master 2 TAL-IL de l'université Paris Nanterre vous devrez proposer un outil qui permette d'extraire automatiquement des réalisations linguistiques à partir de motifs séquentiels émergents, plus largement vous analyserez les résultats et réfléchirez à une mesure d'évaluation pour ces derniers.

1 Contexte sociétal et scientifique

Toute production langagière est évaluée par l'interlocuteur. Il la situe dans un registre, c'est à dire une certaine actualisation de la langue. Les registres de langue sont une notion à l'intersection de la linguistique et de la sociolinguistique. Ils soulèvent une difficulté définitoire en linguistique, sociolinguistique et en traitement automatique des langues.

En linguistique, d'après (Todorov 2013) cette notion s'entend généralement comme renvoyant à 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. De plus, ils se caractérisent par des patrons spécifiques (Ferguson 1982, Ledegen and Léglise 2013).

De fait, l'espace linguistique peut être partitionné en registres qui vont varier selon l'angle d'étude. Par exemple, (Ilmola 2012) privilégie les registres familier,

populaire et vulgaire dans des journaux satiriques, là où (Borzeix and Fraenkel 2005) catégorisent différentes situations de communication au travail.

En linguistique outillée, les travaux anglosaxons menés par (Biber 1991) caractérisent des "registres" qui sont des types textuels générés par des situations de communications différentes. A partir de descripteurs linguistiques listés a priori il propose des études quantitatives de corpus selon différents axes d'analyse : oral/écrit, formel/informel... Son but est d'identifier des descripteurs qui cooccureraient selon ses différents types de textes.

Toutefois, à notre connaissance peu de travaux en français ont étudié les registres de langue en TAL, voire aucun sous l'angle que nous prenons. Cependant de nombreux domaines connexes peuvent être cités tels que l'attribution d'auteurs (Stamatatos 2009, De Vel et al. 2001, Sanderson and Guenter 2006, Koppel and Schler 2003, Argamon et al. 2007), ou bien la classification en sous-genre de textes littéraires : genre romanesque, poétique, épistolaire... (Quiniou et al. 2012, Legallois, Charnois, and Poibeau 2016).

Dès lors nous constatons une difficulté définitoire et terminologique en linguistique et une absence de travaux sur les "registres de langue" dans la littérature scientifique francophone en TAL. C'est pourquoi nous proposons de travailler sur les "registres de langue" en utilisant des outils de TAL afin de dégager des connaissances linguistiques sur ces derniers.

2 Présentation du projet

Le projet TREMoLo a pour "objectifs (...) de progresser dans l'étude des registres de langue et de développer des méthodes automatiques de transformation de textes d'un registre vers un autre."[1]. Pour ce faire nous proposons une chaîne de traitement qui permet d'attribuer automatiquement des proportions de registres aux textes d'un large corpus d'écrits numériques. Ce large corpus est par la suite utilisé pour faire de la fouille de motifs séquentiels émergents, c'est à dire extraire automatiquement des motifs linguistiques spécifiques d'un $registre_1$ par rapport à un $registre_2$.

Une des difficultés des motifs séquentiels émergents réside dans le nombre de motifs retournés et leurs faibles interprétabilités. Aucune évaluation automatique n'a été jusqu'ici proposée et une analyse qualitative humaine est toujours nécessaire. Afin de répondre à ce besoin le projet vise à proposer un outil de correspondance ainsi que des mesures d'évaluation automatique des motifs séquentiels émergents (exemples donnés figure 1).

Le corpus sur lequel vous travaillerez se compose de 268 747 tweets. Il est déjà annoté en lemme, morphosyntaxe et syntaxe.

[1] https://tremolo.irisa.fr/fr/

	Motif	Correspondances
	Familier vs. soutenu	
1	⟨ (*pos:auxiliaire*) , (*syntax:advmod*, *pos:adverbe*, *lemme:pas*) ⟩	• *"Hé ! dis, vieux, je l'ai pas refroidie, au moins ?"* • *"c'est pas non plus ton frometon à toi, béby !"*
2	⟨ (*lemme:c*) , (*pos:ponctuation*, *'*, *lemme:"*, *syntax:ponctuation*) , (*lemme:etre*, *syntax:cop*) ⟩	• *"c'est pas reluisant"* • *"c'est chié la vie avec toi !"* • *"Pfffff. C'était même pas vraix."*
3	⟨ (*pos:ponctuation*, *syntax:ponctuation*) , (*pos:ponctuation*) , (*pos:ponctuation*) ⟩	• *"Et c'est 80 euros d'ailleurs (... ahahahaha)"* • *"ne le laissont pas filer ! ! !"*
4	⟨ (*syntax:nsubj*, *lemme:on*) ⟩	• *"on l'a jamais vu s'afficher avec des meufs"*
5	⟨ (*pos:poncutation*, *mot:?*, *lemme:?*) ⟩	• *"ça compense un manque ou quoi ?"*
6	⟨ (*pos:pronom*, *mot:se*) , (*pos:verbe*) ⟩	• *"pour pas se faire chopper"*
7	⟨ (*pos:pronom_personnel*, *syntaxe:expression_multimots*) ⟩	• *"le Tombeur de Saint-Cloud"* • *"miss Zouzou"*
8	⟨ (*syntax:auxiliaire*) , (*pos:adverbe*) ⟩	• *"C'est bien. Ouais."*
9	⟨ (*pos:verbe*) , (*pos:adverbe*, *syntaxe:modifieur*) , (*pos:adverbe*) ⟩	• *"ça se passera très bien"* • *"où ça se finit pas hyper bien"*
	Soutenu vs. familier	
10	⟨ (*lemme:ne*, *pos:adverbe*) , (*pos:verbe*) ⟩	• *"ne valait-il pas mieux"*
11	⟨ (*pos:pronom*, *mot:me*, *lemme:me*) ⟩	• *"il me semblait"*
12	⟨ (*pos:adverbe*, *mot:vous*, *lemme:vous*) ⟩	• *"vous qui l'aimiez tant"*
13	⟨ (*pos:ponctuation*, *mot:;* , *lemme:;*) ⟩	• *"du Venezuela et du Panamá ; enfin, le Brésil"*
14	⟨ (*mot:comme*, *lemme:comme*) ⟩	• *"comme elle n'avait guère"*

Figure 1: Exemples de motifs séquentiels émergents et de leurs correspondances

3 Grandes étapes du projet

Afin de mener à bien ce projet vous devrez proposer (1) un outil de correspondance entre des réalisations dans les textes du corpus et des patrons morphosyntaxiques retournés par les motifs séquentiels émergents, (2) analyser la qualité des motifs à l'aide d'outils statistiques et en vous basant sur la notion d'émergence, (3) réfléchir à d'autres mesures d'évaluation qui prennent en compte notamment la notion de rang.

4 Livrables

Ce projet doit permettre de livrer :

1. l'ensemble des ressources en python pour extraire automatiquement les motifs séquentiels émergents,
2. les analyses statistiques réalisées,
3. une proposition de métrique d'évaluation des motifs séquentiels émergents.

References

Argamon, Shlomo et al. (2007). "Mining the blogosphere: Age, gender and the varieties of self-expression". In: *First Monday* 12.9.

Biber, Douglas (1991). *Variation across speech and writing*. Cambridge University Press.

Borzeix, Anni and Béatrice Fraenkel (2005). "Langage et travail(communication, cognition, action)". In: *CNRS communication*.

De Vel, Olivier et al. (2001). "Mining e-mail content for author identification forensics". In: *ACM Sigmod Record* 30.4, pp. 55–64.

Ferguson, Charles A (1982). "Simplified registers and linguistic theory". In: *Exceptional language and linguistics*, pp. 49–66.

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

Koppel, Moshe and Jonathan Schler (2003). "Exploiting stylistic idiosyncrasies for authorship attribution". In: *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*. Vol. 69, pp. 72–80.

Ledegen, Gudrun and Isabelle Léglise (2013). *Variations et changements linguistiques*.

Legallois, Dominique, Thierry Charnois, and Thierry Poibeau (2016). "Repérer les clichés dans les romans sentimentaux grâce à la méthode des motifs". In: *Lidil. Revue de linguistique et de didactique des langues* 53, pp. 95–117.

Quiniou, Solen et al. (2012). "Fouille de données pour la stylistique: cas des motifs séquentiels émergents". In: *Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)*, pp. 821–833.

Sanderson, Conrad and Simon Guenter (2006). "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 482–491.

Stamatatos, Efstathios (2009). "A survey of modern authorship attribution methods". In: *Journal of the American Society for information Science and Technology* 60.3, pp. 538–556.

Todorov, Tzvetan (2013). *Mikhaıl Bakhtine. Le principe dialogique. Suivi de: Ecrits du Cercle de Bakhtine*. Le Seuil.