Reconnaissance de défigements dans des tweets en français par similarité d'alignements textuels

Julien Bezançon, *julienbezancon@gmail.com*

11/05/2023

Sorbonne Université

Plan

Introduction

Qu'est-ce qu'un défigement ?

« *toute atteinte à la fixité formelle et à la globalité sémantique d'une séquence figée serait considérée comme un défigement* » [Mejri, 2009]

Qu'est-ce qu'un défigement ?

« *toute atteinte à la fixité formelle et à la globalité sémantique d'une séquence figée serait considérée comme un défigement* » [Mejri, 2009]

Dans l'espace, personne ne vous entendra crier

Qu'est-ce qu'un défigement ?

« *toute atteinte à la fixité formelle et à la globalité sémantique d'une séquence figée serait considérée comme un défigement* » [Mejri, 2009]

Dans l'espace, personne ne vous entendra crier

- Personne ne vous entendra crier dans l'espace

Qu'est-ce qu'un défigement ?

« *toute atteinte à la fixité formelle et à la globalité sémantique d'une séquence figée serait considérée comme un défigement* » [Mejri, 2009]

Dans l'espace, personne ne vous entendra crier

- Personne ne vous entendra crier dans l'espace
- Dans l'**hairspace**, personne ne vous entendra crier

Qu'est-ce qu'un défigement ?

« *toute atteinte à la fixité formelle et à la globalité sémantique d'une séquence figée serait considérée comme un défigement* » [Mejri, 2009]

Dans l'espace, personne ne vous entendra crier

- Personne ne vous entendra crier dans l'espace
- Dans l'**hairspace**, personne ne vous entendra crier
- **En calbute à la maison**, personne ne vous entendra crier

Création du corpus

Quelques chiffres

- 3 362 750 tweets extraits (novembre 2020 - janvier 2023)

Quelques chiffres

- 3 362 750 tweets extraits (novembre 2020 - janvier 2023)
- 99 244 tweets après le premier filtrage

Quelques chiffres

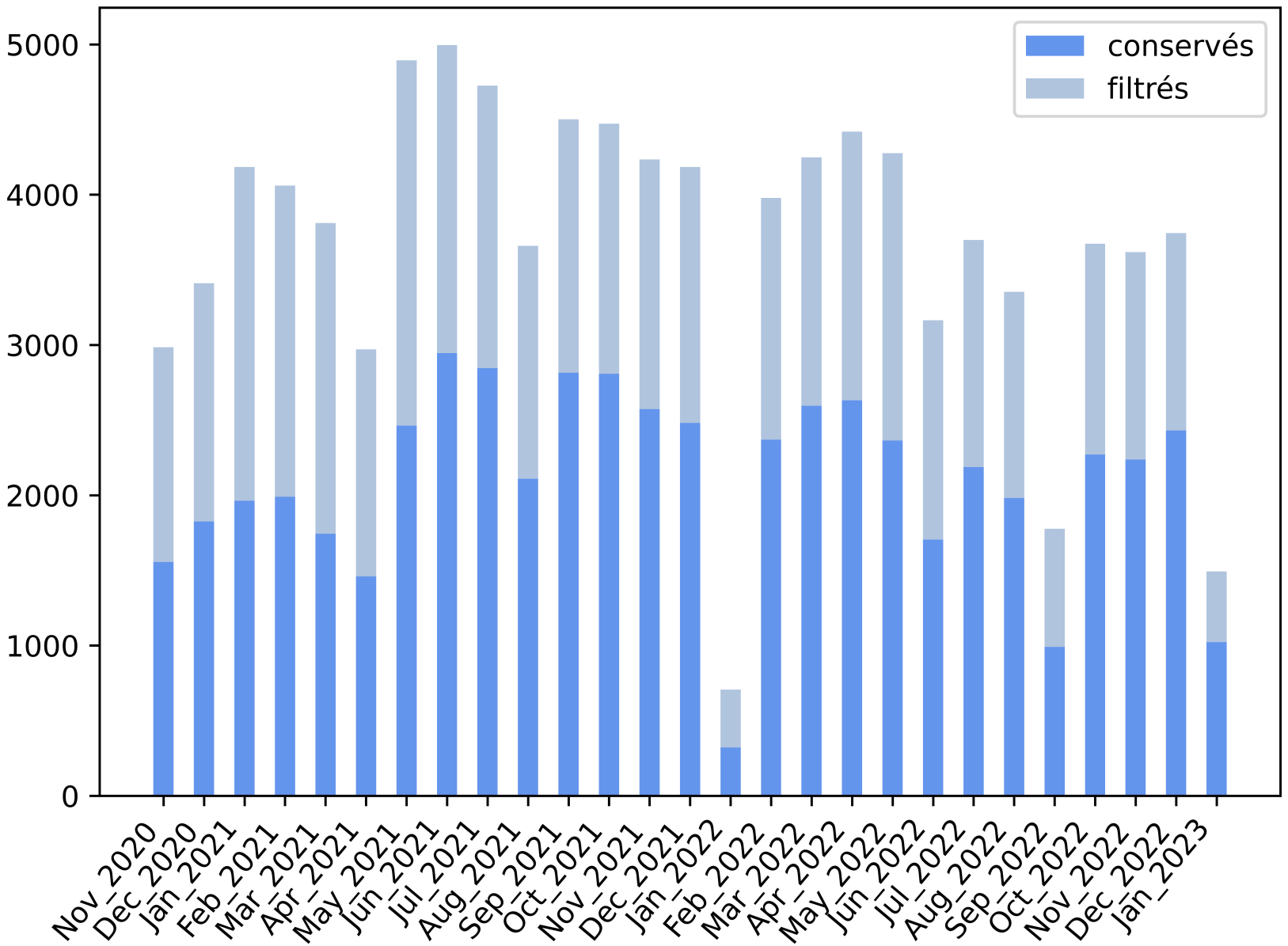
- 3 362 750 tweets extraits (novembre 2020 - janvier 2023)
- 99 244 tweets après le premier filtrage
- 56 687 tweets après le second filtrage

Quelques chiffres

- 3 362 750 tweets extraits (novembre 2020 - janvier 2023)
- 99 244 tweets après le premier filtrage
- 56 687 tweets après le second filtrage

Nous ne conservons que 1,68 % des tweets extraits au début !

Quelques chiffres



Détection des défigements candidats

Alignements de séquences

« *l'alignement de séquences est une manière de représenter deux ou plusieurs séquences [...] les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires* » (Wikipédia)

Travailler	plus	pour	gagner	-	plus
Travailler	plus	pour	gagner	**moins**	-

Alignements de séquences

« *l'alignement de séquences est une manière de représenter deux ou plusieurs séquences [...] les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires* » (Wikipédia)

Travailler	plus	pour	gagner	-	plus
Travailler	plus	pour	gagner	**moins**	-

quelques français les suivent pensant **travailler encore moins pour gagner encore plus**

Analyse sur plusieurs couches linguistiques

Couches	Expression	Segment du tweet capturé					Sim
Brute	Que la force soit avec toi	**Que**	**la**	*(et la chance)*	**soit**	**avec**	0.76
		force			**toi**		
Lemmatisée	Que le force être avec toi	**Que**	**le**	*(et le chance)*	**être**	**avec**	0.62
		force			**toi**		
Étiquetée	SCONJ DET NOUN VERB ADP PRON	*PRON*	**DET**	*PUNCT CCONJ*	*CCONJ*	**ADP**	0.36
		NOUN		*DET NOUN PUNCT*	**PRON**		
Phonétisée	kə la fɔrs swa avək twa	kə la fɔrs		(ɛ la ʃãsə)	swa	avək	0.76
					twa		

Quelques exemples de défigements candidats

Défigements en milieu d'énoncé

- **Dans la Beauce** personne ne vous entendra crier
- Travailler plus pour **redistribuer** plus

Quelques exemples de défigements candidats

Défigements en milieu d'énoncé

- **Dans la Beauce** personne ne vous entendra crier
- Travailler plus pour **redistribuer** plus

Défigements en fin d'énoncé

- Dans l'espace, personne ne vous entendra crier **BONNE ANNÉE**
- Travailler plus pour **ne plus rien gagner**

Quelques exemples de défigements candidats

Défigements en milieu d'énoncé

- **Dans la Beauce** personne ne vous entendra crier
- Travailler plus pour **redistribuer** plus

Défigements en fin d'énoncé

- Dans l'espace, personne ne vous entendra crier **BONNE ANNÉE**
- Travailler plus pour **ne plus rien gagner**

Défigements ouverts sur leur contexte droit

- Ce moment où **tu prends conscience que tu ne mérites pas ça**
- Je traverse la rue et **je te trouve un boulot**

Tri par mesures de similarités

Échantillon de défigements candidats capturés

Défigement candidat	Cos	Dic	Ham	Jac	Kul	Mat	Rus	Freq
travailler plus pour gagner moins	0,8	0,82	0,6	0,7	0,54	0,7	0,7	364
travailler plus pour gagner pareil	0,8	0,82	0,6	0,7	0,54	0,7	0,7	10
travailler plus pour gagner autant	0,8	0,82	0,6	0,7	0,54	0,7	0,7	9
travailler plus pour travailler plus	0,78	0,71	0,33	0,56	0,38	0,56	0,56	5
travailler plus pour gagner un peu plus	0,78	0,7	0,54	0,54	0,37	0,54	0,54	3
travailler plus pour payer plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	44
travailler plus sans gagner plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	21
travailler plus pour partager plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	10
travailler plus pour produire plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	9
travailler plus et gagner plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	9
travailler plus pour être plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	7
travailler plus pour mourir plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	7
travailler plus pour crever plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	7
travailler plus pour donner plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	6
travailler plus pour perdre plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	5
travailler plus pour avoir plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	4
travailler plus pour faire plus	0,73	0,62	0,45	0,45	0,29	0,45	0,45	4
travailler moins pour gagner plus	0,7	0,71	0,45	0,55	0,38	0,55	0,55	176
travailler autant pour gagner plus	0,7	0,71	0,45	0,55	0,38	0,55	0,55	5
travailler à l étranger pour gagner plus	0,71	0,7	0,45	0,55	0,38	0,55	0,55	2

Les exceptions

Coquilles et autres fautes

- travailler plus poyr gagner plus
- travailler plus pour gagné plus
- travailler plus pour gagniez plus
- travailler plus pour gagnez plus

Captures incomplètes

- travailler plus pour le plus
- travailler plus pour la plus
- travailler plus pour l ukraine plus
- travailler plus pour être plus

Conclusion

Conclusion

Étapes réalisées

- Création d'un corpus
- Extraction de défigements candidats
- Tri des défigements candidats par mesure de similarité

Perspectives pour la suite

- Amélioration de la capture
- Analyse du tri effectué : robuste ? pertinent ?
- « Faux » défigements candidats : comment les retirer ?

Références


Eline, J. and Zhu, L. (2014).
Défigement et inférence - cas d'études du Canard enchaîné.
SHS Web of Conferences, 8 :681 695.


Lamiroy, B. (2008).
Le figement : à la recherche d'une définition.
ZFSL, Zeitschrift für französische Sprache und Literatur, 36 :85–99.


Mejri, S. (2009).
Figement, défigement et traduction. Problématique théorique.
Pratiques, page 153.