Reconnaissance de défigements dans des tweets en français par des mesures de similarité sur des alignements textuels

Julien Bezançon, Gaël Lejeune

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

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

RÉSUMÉ

Cet article propose une première approche permettant la reconnaissance automatique de défigements linguistiques dans un corpus de tweets. Les recherches portant sur le domaine du figement ont gagné en popularité depuis quelques décennies. De nombreux travaux dérivés de cette notion sont également apparus, portant sur le phénomène corollaire du défigement. Alors que les linguistes essayent de décrypter les modes de construction de ces exemples de créativité lexicale, peu de travaux de recherche en TAL s'y sont intéressés. La problématique qu'offre le cas du défigement est pourtant intéressante : des outils informatiques peuvent-ils être en mesure de reconnaître automatiquement un défigement ? Nous présentons ici une méthodologie basée sur des alignements de séquences réalisés sur diverses couches d'informations linguistiques. Cette méthodologie permet l'isolement de potentiels défigements au sein d'un corpus de tweets. Nous expérimentons ensuite une méthode de tri par similarité des défigements potentiels isolés.

ABSTRACT

Recognition of unfrozen expressions in french tweets using similarity measures

This paper proposes a first approach for the automatic recognition of unfrozen expressions in a corpus of tweets. Research on frozen expressions has been gaining in popularity for a few decades. Similarly, many works derived from this notion have emerged, dealing with the phenomenon of unfreezing. While linguists try to understand the modes of appearance of this phenomenon and its relation to the freezing effect, no research work in computer science has focused on it. However, the scientific question that arises with freezing/unfreezing is interesting : can computer tools automatically recognize an unfrozen expression ? We present here a methodology based on sequence alignments performed on various linguistic layers. This methodology allows the isolation of possible unfrozen expressions within a corpus of tweets. We then use different similarity methods to sort these possible unfrozen expressions.

MOTS-CLÉS : Figement linguistique, Expression Figées, Défigement, Alignement, Similarité.

KEYWORDS: Frozen expressions, Unfrozen expressions, Alignment, Similarity.

1 Introduction

Le figement est le phénomène par lequel une séquence de mots va se fixer, faisant ainsi perdre aux différents termes leur sens individuel au profit d'un sens global faiblement compositionnel. Cette notion figure au cœur de nombreux travaux, bien qu'il n'y ait pas de définition univoque de ce phénomène (Lamiroy, 2008). Il en résulte un manque de propriétés formelles fiables permettant la reconnaissance automatique de ces phénomènes. Toutefois, certains chercheurs ont développé des ressources lexicales et des méthodes consacrées à la reconnaissance automatique de séquences figées (ci-après SF) en français (Leclère, 2000; Mejri, 2005; Fort *et al.*, 2018, 2020) comme pour d'autres langues (Baptista *et al.*, 2004; Català & Baptista, 2007; Tan *et al.*, 2021).

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 56–67.

La notion de défigement linguistique est intimement liée à la notion de figement. Le défigement est un phénomène qui vient briser une SF en lui retirant son caractère monolithique à la suite d'une ou plusieurs transformations linguistiques (Eline & Zhu, 2014). (Gross, 1996) va jusqu'à considérer le défigement comme un critère de reconnaissance du figement. À notre connaissance, contrairement aux SF, il n'existe pas de travaux de recherches s'intéressant spécifiquement à la reconnaissance automatique de séquences défigées (SD) à l'exception peut être de travaux sur la production de jeu de mots, énoncés que l'on peut voir comme des cas particuliers de défigements linguistiques (Valitutti *et al.*, 2013). Pourtant, ceci serait d'un intérêt triple d'un point de vue linguistique : (I) caractériser le figement d'expressions par leur productivité en termes de SD ; (II) détecter l'apparition de figements en « en temps réel » et (III) étudier les processus permettant aux locuteurs humains de reconnaître ces défigements. Nous faisons l'hypothèse qu'il est possible d'exploiter une méthodologie automatique permettant à des experts linguistiques d'explorer des corpus de manière non-supervisée afin d'opérer une veille informatisée d'un grand intérêt pour des chercheurs en linguistique.

Cet article est organisé de la façon suivante : dans la section 2 nous étudions les notions de figement et de défigement afin d'identifier les propriétés permettant de reconnaître automatiquement ces défigements à l'aide d'outils informatiques, dans la section 3 nous composons un corpus de tweets avec des défigements candidats, puis dans les sections 4 et 5 nous proposons des observations qualitatives et quantitatives sur les séquences défigées que nous avons extraites avant de proposer quelques perspectives de ce travail dans la section 6.

2 Propriétés linguistiques du défigement

Intrication des concepts de figement et de défigement Le figement linguistique se définit par trois critères principaux, issus de l'abondante littérature sur le sujet (Gross, 1982; Rommers *et al.*, 2013; Molinaro & Carreiras, 2010; Lamiroy, 2008; Mejri, 2005) :

non-compositionnalité du sens : on ne peut pas réduire le sens d'un figement au sens de chaque unité le composant, un sens global émerge ;

non-modifiabilité de la structure : l'ordre des mots composant le figement reste identique ;

non-substituabilité des termes : on ne peut pas remplacer l'un des termes du figement, même par un synonyme.

Le défigement vient « briser » un figement, les trois critères décrivant le figement permettent donc de détecter qu'un défigement se manifeste : « toute atteinte à la fixité formelle et à la globalité sémantique d'une SF serait considérée comme un défigement » (Mejri, 2009). Les études sur la notion de défigement s'accordent à dire que la reconnaissance d'un défigement implique au préalable la reconnaissance par le locuteur du figement dont il est issu (Fiala & Habert, 1989; Eline & Zhu, 2014) et que seul ce qui est figé se défige (Greciano, 1985). Pour caractériser chaque défigement, il faut par conséquent pouvoir identifier le figement à partir duquel ce défigement se forme. Une première piste pour notre travail de reconnaissance de SD est de créer une ressource de SF afin d'orienter les recherches et guider la reconnaissance. Selon (Eline & Zhu, 2014), au contraire du figement, le défigement est motivé, on peut donc considérer que l'on a une « remotivation » du figement, il faut que sa formation soit justifiée. Ce critère permet notamment de différencier l'énoncé fautif, par exemple une SF mal reconstituée par un locuteur [1], d'un réel défigement. Ce critère étant difficile à caractériser automatiquement, nous considérons ici que cette distinction serait réalisée par un expert linguiste en aval de la chaîne de traitement automatique. (Eline & Zhu, 2014) précisent également

1. Par exemple dire « être sur un même pied d'égalité » au lieu de « être sur un pied d'égalité »

que « toutes les séquences figées sont défigeables, mais toutes n'ont pas la même possibilité de défigement ». Les auteurs ajoutent que « plus la construction d'une SF est rare, plus la SF est apte à un défigement. ». Ces propriétés de productivité en défigements d'une SF et de fréquence nous seront utiles pour pouvoir mesurer parmi les défigements candidats ceux qui sont les plus intéressants en termes de créativité lexicale.

Degré de figement et défigement Toute séquence de mots possible dans la langue est en réalité plus ou moins figée, se plaçant dans un *continuum* allant de l'énoncé libre à la séquence figée (Gross, 1982; Mejri, 1998). Pour décrire ce phénomène, la littérature fait référence aux notions de degré de figement et d'opacité sémantique. Les semi-figements et quasi-figements (Mejri, 2005; François & Manguin, 2006) s'opposent ainsi aux figements absolus qui sont des figements non formellement contraints. Ces catégories de figement acceptent différentes variations et les SF y appartenant sont identifiées grâce à différents tests linguistiques tels que ceux effectués dans (Gross, 1982). (Nunberg *et al.*, 1994) précise qu'une SF peut posséder une interprétation littérale et une interprétation figée, pouvant être corrélées au sens compositionnel. En plus de variations formelles, nous pouvons donc observer des variations sémiques et un sens plus ou moins global en fonction des SF. Cette notion de degré de figement peut être problématique pour le TAL : une perte de la globalité du sens signifie-t-elle que nous sommes face à un défigement, ou à une SF avec un faible degré de figement ? Comment différencier une SF avec un degré de figement faible d'une SD ? Existe-t-il des cas ambigus ? Nous observons une relation entre degré de figement et défigement. Selon (Cusimano, 2015), une expression figée avec un degré de figement élevé a plus de chance de donner un défigement avec un degré de figement élevé.

Procédés de formation des défigements Nous identifions plusieurs procédés par lesquels des SD sont formées. (Galisson, 1993) parle de filiation phonique et de destructuration syntaxique. (Lecler, 2004) considère les défigement comme marqueurs dialogiques. (Eline & Zhu, 2014) font état d'un viol de la structure syntaxique, de la norme orthographique et d'une détérioration de la structure formelle des figements. Afin de visualiser ces changements avec des outils informatiques, nous devons récupérer des informations linguistiques sur plusieurs couches. Nous désirons analyser les couches syntaxiques, phonétiques et lexicales des expressions figées de notre corpus de figement et des possibles défigements de notre jeu de données. Tous les procédés de formation des défigements que nous venons de décrire illustrent des défigements marqués formellement : nous pouvons distinguer à la lecture le figement dont ils sont issus et les modifications apportées à la SF. Nous devons les dissocier des défigements non marqués formellement (Eline & Zhu, 2014). Ces défigements ne connaissent pas de changements formels par rapport aux figements dont ils sont issus. Leur statut de défigement est dû à des modifications de sens ou de prononciation par rapport à leur figement d'origine. Ils s'identifient donc exclusivement à l'aide du contexte dans lequel on les retrouve. Pour cette étude, nous nous intéressons exclusivement aux défigements analysables hors contexte. Afin de disposer d'un corpus de taille suffisante avec une créativité lexicale variée, nous avons choisi de nous intéresser aux réseaux sociaux numériques, et en l'espèce à TWITTER. Il s'agit de disposer de suffisamment d'exemples originaux pour avoir à la fois des « vrais positifs », de réels défigements, mais aussi des faux positifs, des contre-exemples qui vont permettre de questionner la qualité des algorithmes d'identification. La création de ce corpus est décrite dans la section suivante.

3 Identification de défigements candidats dans des tweets

Le processus que nous avons construit comporte trois étapes que nous décrivons ici : (I) l'extraction de tweets sources à partir de SF connues et le filtrage de ces tweets, (II) l'alignement des SF et des tweets et (III) l'isolement des segments communs constituant des défigements potentiels.

Extraction de tweets et filtrage. Pour collecter nos tweets, nous avons pris une base témoin de 217 expressions figées [2] appartenant à quatre catégories : (1) extraits ou slogans de publicité ; (2) citations politiques ou historiques ; (3) accroches pour des films de cinéma et (4) autres types de locutions. Nous avons cherché autant que possible à répartir équitablement les SF de ces catégories entre des expressions anciennes et plus récentes d'une part, et des expressions dont la connaissance est *a priori* répandue chez les locuteurs du français, ou au contraire plus confidentielle d'autre part. Ce critère est quelque peu subjectif, mais l'objectif est de pouvoir regarder en détails un nombre limité d'expressions. Nous donnons ci-dessous un exemple de SF pour chaque catégorie :

1. Tu pousses le bouchon un peu trop loin, Maurice.
2. On ne peut pas accueillir toute la misère du monde.
3. Dans l'espace personne ne vous entend crier.
4. Partir, c'est mourir un peu.

Les mots des SF sélectionnées sont utilisés pour faire des requêtes via l'API de TWITTER. Pour chaque requête, nous donnons une SF complète, ceci afin de ne pas trop biaiser le corpus d'étude en contraignant trop les résultats. Nous avons réalisé 3 collectes quotidiennes entre novembre 2020 et janvier 2023 inclus aboutissant à un total de 3 362 750 tweets extraits. Chaque tweet est associé à la SF qui nous a permis de l'extraire. Nous avons ensuite pré-filtré les tweets en ne conservant que ceux qui comportent au moins 50 % de mots en commun avec la SF d'origine. Au final, nous obtenons 99 244 tweets. Nous excluons ensuite les « figements » , c'est-à-dire les énoncés qui contiennent strictement l'expression figée recherchée. Le résultat de ce nouveau filtrage est représenté dans la Figure 1. Il résulte de tout cela un total de 56 687 tweets contenant potentiellement des défigements et 42 557 tweets contenant simplement un figement.

Encodage et alignement des séquences figées et de leurs défigements potentiels. Pour chaque tweet de notre corpus, nous calculons des alignements séquentiels pour visualiser les différences entre un défigement candidat et la SF associée. Ces alignements sont basés sur un découpage des tweets en mots et une analyse sur différentes couches d'informations linguistiques :

brute : tokenisation des formes de base du tweet et de la séquence figée ;

lemmatisée : recherche des formes canoniques pour traiter les micro-variations formelles ;

étiquetée syntaxiquement : pour rechercher la proximité de structure avec la SF ;

phonétisée : encodage en phonèmes pour valoriser les SD jouant sur la sonorité.

Afin de faciliter un futur traitement multilingue, nous avons utilisé SPACY pour les trois premières couches (tokenisation, lemmatisation et étiquetage) et EPITRAN [3] pour la phonétisation. Ensuite, nous alignons la SF et le tweet à l'aide de BIOPYTHON [4]. Cette librairie permet de calculer tous les alignements possibles entre deux séquences avec un alignement au token. Prenons l'expression

2. la liste complète est donnée sur le dépôt GitHub de notre projet https://github.com/JulienBez/DefigementTALN2023
3. https://pypi.org/project/epitran/
4. https://biopython.org/

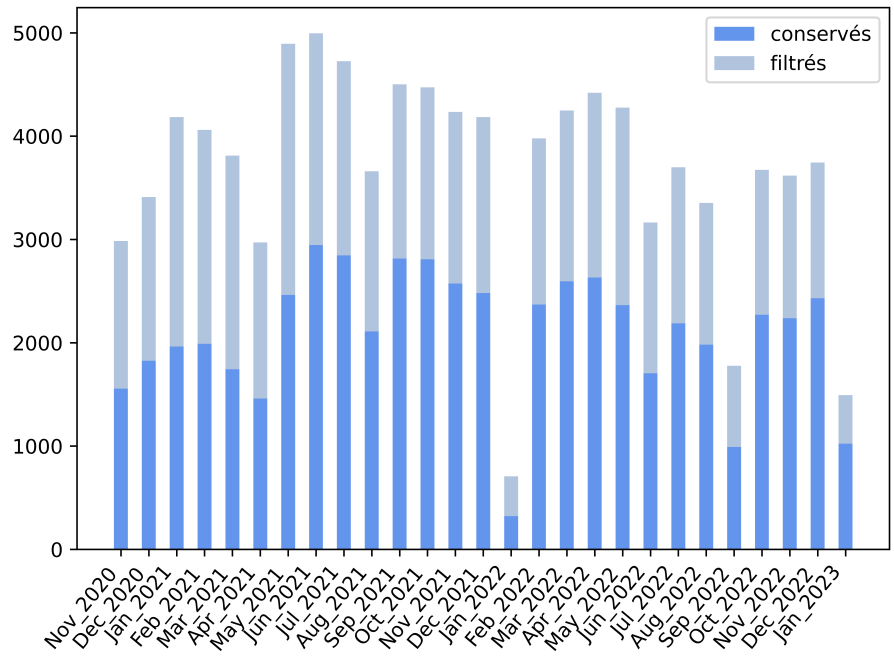

FIGURE 1 – Résultats de la récolte de tweets de novembre 2020 à janvier 2023 avant et après filtrage.

figée « Travailler plus pour gagner plus » et le défigement suivant, fréquent (voire banal) dans notre corpus : « Travailler plus pour gagner moins », l'alignement séquentiel obtenu serait :

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

Isolement des potentiels défigements. Les alignements contribuent à l'isolement d'une SD au sein d'un tweet, mais l'isolement n'est pas totalement réalisé puisque le tweet où le défigement est idéntifié peut contenir beaucoup plus de tokens que la SF d'origine. Les alignements sont commodes pour la lecture rapprochée des résultats mais ne permettent pas pour autant de délimiter précisément un défigement dans un tweet. L'étape suivante consiste donc à extraire les segments communs entre chaque tweet et son expression figée, c'est-à-dire d'extraire la plus grande séquence de mots commençant par le premier terme trouvé dans le tweet appartenant à l'expression figée et finissant par le dernier terme trouvé dans le tweet appartenant à cette expression. L'exemple ci-dessous représente un tweet de notre jeu de données, la SF qui a permis de sélectionner ce tweet et le segment commun identifié avec une distance d'édition au token de 2 (marqués en gras).

Tweet :	@user1 @user2 C bon ça ! !	Travailler	**moins**	pour	**bronzer**	plus	bye
SF :		Travailler	**plus**	pour	**gagner**	plus	
Défigement :		Travailler	**moins**	pour	**bronzer**	plus	

Les segments communs sont formés à partir de chaque alignement disponible pour un tweet (à raison d'un segment commun par alignement). Nous obtenons au moins un segment commun pour chaque couche d'information extraite du tweet (brute, lemmatisée, étiquetée et phonétisée). Dans le cas où, pour une même couche linguistique, nous avons plusieurs alignements, nous les conservons tous. Nous choisissons parmi ces alignements celui dont le segment commun est le plus proche de la SF [5]. Le Tableau 1 illustre de manière simplifiée ce que nous comparons entre une expression figée et le

5. Sans être strictement identique, puisque nous ne travaillons que sur les tweets ne comprenant pas la SF recherchée.

Couche	Séquence						Segment commun dans le tweet			Sim
BRUTE	Que	la	force	soit	avec	toi	**Que la force**	*(et la chance)*	**soit avec toi**	.76
LEMME	Que	le	force	être	avec	toi	**Que le force**	*(et le chance)*	**être avec toi**	.62
POS	SCONJ	DET	NOUN	VERB	ADP	PRON	*PRON* **DET NOUN**	*CCONJ DET NOUN*	*CCONJ* **ADP PRON**	.36
PHON	kə	la	fɔrs	swa	avək	twa	kə la fɔrs	(ɛ la ʃãsə)	swa avək twa	.76

TABLE 1 – Extraction du segment commun entre un tweet et une expression pour chaque couche linguistique (en gras les sous-séquences communes, en italiques les sous-séquences insérée).

segment commun extrait d'un tweet, nous y ajoutons la similarité cosinus calculée sur les vecteurs de bigrammes d'unités pour chaque couche (tokens, lemmes ...). Le segment commun permet de réduire le contenu du tweet à la sous-séquence constituant un potentiel défigement et donc de mieux mesurer la similarité que si nous gardions l'intégralité du tweet pour calcul. Pour les défigements simples (substitution ou insertion), le segment commun peut correspondre intégralement au potentiel défigement recherché, comme c'est le cas dans le Tableau 1.

4 Observations sur la qualité des défigements candidats

Cas de l'absence de défigement. Sans surprise, nous isolons des segments communs ne correspondant pas du tout à des défigements. Il s'agit de séquences dites libres. (1) et (2) en sont des exemples.

1. Tout ça pour travailler samedi prochain en plus yes. (SF : « travailler plus pour gagner plus »)
2. En train de regarder le plus beau travailler. (SF : « train de vie »)

Expressions sans trace de défigement. Nous n'utilisons pas les tweets contenant exactement la SF. Ces tweets ne font pas état d'un quelconque défigement. Nous notons deux situations ambiguës liées à la non prise en compte des tweets contenant l'expression recherchée mot pour mot : les défigements qui se prolongent dans le contexte du figement et les défigements non marqués formellement.

Défigement en milieu d'énoncé C'est le type de défigement le mieux identifié par notre méthodologie. Voici quelques exemples identifiés à partir des SF « Dans l'espace, personne ne vous entendra crier » et « Travailler plus pour gagner plus ».

3. Dans **la Beauce personne** ne vous entendra crier.
4. Travailler plus pour **redistribuer** plus.

Défigement en début/fin d'énoncé. Parfois, les modifications se situent en début ou fin d'expression, il est alors plus difficile de borner le segment commun. Ainsi, notre méthode ne permet pas d'isoler :

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

Expressions ouvertes sur leur contexte droit. Elles ont la spécificité d'être ouvertes sur leur contexte droit. Nous avons ainsi pour les SF « Ce moment où ... » et « Je traverse la rue et ... » :

7. Ce moment où **tu prends conscience que tu ne mérites pas ça...**
8. je traverse la rue et **je te trouve un boulot**.
9. Moi je traverse la rue et **je t'en gagne une de médaille d'or**.

Pour le moment, nous n'appliquons pas de traitements particuliers à ce type d'expressions. Nous envisageons cependant de préciser que pour ces expressions, il nous faut récupérer tout le contexte droit jusqu'à la première marque de fin de phrase (un point) ou même jusqu'à la fin du tweet.

5 Tri des défigements candidats par mesures de similarité

Afin d'évaluer si une approche par mesures de similarités permet de détecter des défigements, nous créons plusieurs classements des segments communs. Un classement est créé pour chaque mesure de similarité et pour chaque couche d'information linguistique. Nous cherchons à comparer systématiquement l'expression figée recherchée et les segments communs correspondants. Nous supposons qu'il doit exister un seuil, variable selon les expressions et les couches, à partir duquel les résultats renvoyés correspondent régulièrement à des SD. Nous testons différentes mesures de similarités (Cosinus, Dice, Hamming, Jaccard, Kulsinski, Matching et Rusell-rao). Nous allons observer plus en détails les défigements obtenus par couche linguistique pour la SF « Travailler plus pour gagner plus » (parmi 7 520 tweets contenant des défigements candidats). Nous prenons les dix défigements candidats les plus fréquents et nous les classons par mesure de similarité pour chaque couche. Les résultats obtenus avec la mesure de similarité cosinus sont présentés dans les Tableaux 2 à 5. Une SD remarquable a tendance à être plus utilisée selon (Cusimano, 2015), d'où notre choix de travailler sur les 10 candidats les plus fréquents.

Défigement potentiel	Sim	Freq
travailler plus pour gagner moins	0,80	364
travailler plus pour gagner pareil	0,80	10
travailler plus pour gagner autant	0,80	9
travailler plus pour payer plus	0,73	44
travailler plus sans gagner plus	0,73	21
travailler plus pour partager plus	0,73	10
travailler moins pour gagner plus	0,70	176
travailler plus pour vivre moins	0,60	16
travailler plus pour perdre moins	0,60	14
travailler moins et gagner plus	0,50	30

TABLE 2 – Couche brute

Défigement potentiel	Sim	Freq
travailler plus pour gagner moins	0,80	364
travailler plus pour gagner pareil	0,80	10
travailler plus pour gagner autant	0,80	9
travailler plus pour payer plus	0,73	44
travailler plus sans gagner plus	0,73	21
travailler plus pour partager plus	0,73	10
travailler moins pour gagner plus	0,70	176
travailler plus pour vivre moins	0,60	16
travailler plus pour perdre moins	0,60	14
travailler moins et gagner plus	0,50	30

TABLE 3 – Couche lemmatisée

Défigement potentiel	Sim	Freq
travailler plus pour gagner moins	1,0	438
travailler moins pour gagner plus	1,0	179
travailler plus pour payer plus	1,0	44
travailler plus sans gagner plus	1,0	21
travailler plus pour vivre moins	1,0	18
travailler plus pour perdre moins	1,0	14
travailler plus pour gagner autant	1,0	12
gagner plus sans travailler plus	1,0	11
travailler plus pour produire plus	1,0	10
travailler plus pour mourir plus	1,0	7

TABLE 4 – Couche étiquetée

Défigement potentiel	Sim	Freq
travailler plus pour gagner moins	0,80	364
travailler plus pour gagner pareil	0,80	10
travailler plus pour gagner autant	0,80	9
travailler plus pour payer plus	0,73	44
travailler plus sans gagner plus	0,73	21
travailler plus pour partager plus	0,73	10
travailler moins pour gagner plus	0,70	176
travailler plus pour vivre moins	0,60	16
travailler plus pour perdre moins	0,60	14
travailler moins et gagner plus	0,57	30

TABLE 5 – Couche phonétisée

Nous remarquons pour les couches brute, lemmatisée et phonétisée des résultats identiques avec sept défigements pour des mesures de similarités allant de 0,7 à 0,8 et trois défigements pour des mesures de similarité allant de 0,6 à 0,5. Nous observons deux différences notables entre les résultats obtenus avec ces trois couches et les résultats obtenus à la couche étiquetée. La première est une différence de fréquence des segments communs. Cela peut s'expliquer par le nombre important d'alignements trouvés pour la couche étiquetée (34 008 alignements) par rapport aux autres couches (11 708 pour

les couches brute et lemmatisée, 11 710 pour la couche phonétisée). Nous rappelons que pour chaque tweet, un alignement est renvoyé par possibilité d'alignement identifiée avec la librairie BIOPYTHON et ce pour chaque couche linguistique étudiée. Comme les étiquettes morphosyntaxiques d'une SF peuvent être assez communes et se retrouver de multiples fois dans un tweet, il n'est pas surprenant que les alignements morphosyntaxiques soient multiples. La seconde différence concerne les mesures de similarité obtenues. Sont renvoyés avec une mesure de similarité égale à 1 tous les segments communs dont la structure morphosyntaxique correspond exactement à celle de la SF. Bien que les segments communs renvoyés dans le Tableau 4 correspondent tous à des défigements, nous retrouvons également des segments communs comme (10), (11) et (12) avec une même similarité de 1.

10. Fait tout pour agir maintenant
11. Je fais comment pour travailler
12. Commence alors à décliner progressivement

Pour cette expression, la couche étiquetée ne s'avère pas efficace, par contre en observant les trois autres couches, nous observons qu'un seuil de similarité de 0,7 pour Cosinus et Dice permet de ne garder que des SD. Nous présentons dans le Tableau 6 les 20 résultats les plus fréquents obtenus avec la couche brute et une similarité cosinus entre 0,8 et 0,7 et nous y ajoutons les résultats obtenus avec les autres mesures de similarité pour chaque segment commun. Nous remarquons (lignes grisées) un ensemble de SD possédant les mêmes résultats d'une mesure de similarité à l'autre (0,73 pour la similarité cosinus). Au total, nous comptons 74 potentiels défigements avec les mêmes résultats pour chaque mesure de similarité pour notre SF. Sur ces 73 potentiels défigements, 65 sont des SD et 8 ne sont pas des SD. Nous divisons les séquences qui ne sont pas des SD en deux catégories : les SF avec un énoncé fautif (13 à 16) et les SD qui n'ont été capturées que partiellement (17 à 20).

13. travailler plus poyr gagner plus
14. travailler plus pour gagné plus
15. travailler plus pour gagniez plus
16. travailler plus pour gagnez plus
17. travailler plus pour le plus
18. travailler plus pour la plus
19. travailler plus pour l ukraine plus
20. travailler plus pour être plus

Toutes les SD obtenues pour ces résultats sont globalement de la même forme : il s'agit de défigements par substitution d'un des termes de l'expression recherchée par un autre terme. Nous pourrions isoler un seuil de similarité dans lequel nous observerions toutes les occurrences de SD par substitution d'un terme. Cependant, ce seuil varie selon les expressions. On remarque des SD formées par substitution d'un terme avec des similarités supérieures et inférieures à celles que nous analysons, que nous indiquons en bleu dans le Tableau 6. Nous observons que la position des termes substitués a un impact sur les mesures de similarités renvoyées. Le seuil d'identification varie sans doute selon le procédé de formation du défigement. Notons que ces seuils ne garantissent pas l'absence de faux positifs. Il faudrait pouvoir quantifier ces faux positifs pour chaque seuil et établir une méthode permettant de les trier. Par exemple, pour les fautes d'orthographe, nous pouvons observer la couche lemmatisée : pour notre seuil, on éliminerait ainsi les faux positifs (13), (14) et (15), qui contiennent des formes mal orthographiées du verbe « gagner ». De même, si nous recherchons des défigements par substitution, il est très probable que les SD aient une similarité de 1 sur la couche morphosyntaxique avec nos mesures de similarité. On exclut ainsi (17), (18), (19) et potentiellement (13) du fait de la faute d'orthographe, non répertoriée dans le lexique et non-étiquetée. En revanche, (20) n'est pas écarté avec ces filtrages.

Défigement potentiel					Cos	Dic	Ham	Jac	Kul	Mat	Rus	#
travaille	plus	pour	gagner	moin	,80	,82	,60	,70	,54	,70	,70	364
travaille	plus	pour	gagner	parei	,80	,82	,60	,70	,54	,70	,70	10
travaille	plus	pour	gagner	autar	,80	,82	,60	,70	,54	,70	,70	9
travaille	plus	pour	travailler	plus	,78	,71	,33	,56	,38	,56	,56	5
travailler	plus	pour	gagner un peu	plus	,78	,70	,54	,54	,37	,54	,54	3
travaille	plus	pour	payer	plus	,73	,62	,45	,45	,29	,45	,45	44
travaille	plus	sans	gagner	plus	,73	,62	,45	,45	,29	,45	,45	21
travaille	plus	pour	partager	plus	,73	,62	,45	,45	,29	,45	,45	10
travaille	plus	pour	produire	plus	,73	,62	,45	,45	,29	,45	,45	9
travaille	plus	et	gagner	plus	,73	,62	,45	,45	,29	,45	,45	9
travaille	plus	pour	être	plus	,73	,62	,45	,45	,29	,45	,45	7
travaille	plus	pour	mourir	plus	,73	,62	,45	,45	,29	,45	,45	7
travaille	plus	pour	crever	plus	,73	,62	,45	,45	,29	,45	,45	7
travaille	plus	pour	donner	plus	,73	,62	,45	,45	,29	,45	,45	6
travaille	plus	pour	perdre	plus	,73	,62	,45	,45	,29	,45	,45	5
travaille	plus	pour	avoir	plus	,73	,62	,45	,45	,29	,45	,45	4
travaille	plus	pour	faire	plus	,73	,62	,45	,45	,29	,45	,45	4
travaille	moins	pour	gagner	plus	,70	,71	,45	,55	,38	,55	,55	176
travaille	autant	pour	gagner	plus	,70	,71	,45	,55	,38	,55	,55	5
travailler	à l étranger	pour	gagner	plus	,70	,70	,45	,55	,38	,55	,55	2

TABLE 6 – Résultats obtenus avec la SF « Travailler plus pour gagner plus » pour chaque mesure de similarité, classés par similarité cosinus décroissante.

Nous réalisons la même expérience avec une nouvelle SF : « Que la force soit avec toi » (1 523 occurrences). Nous obtenons le Tableau 7. Nous retrouvons en gris les potentiels défigements dont la mesure de similarité cosinus est égale à 0,73. Là encore, nous retrouvons pour ces résultats des défigements formés par la substitution d'un des termes de la SF. Nous remarquons tout de même des différences entre les résultats des mesures de similarités d'une SF à l'autre (sauf pour les similarités Cosinus et Kulsinski).

Les SD « que la force et le courage soit avec toi » et « que le pouvoir de la force soit avec toi » ont bien une mesure de similarité cosinus égale à 0,73 mais leur procédé de formation n'est pas la substitution d'un des termes de l'expression recherchée. Nous remarquons des résultats différents pour toutes les autres mesures de similarités pour ces deux SD, ce qui permet de les isoler des autres SD avec une distance cosinus de 0,73 qui sont bien formées par substitution d'un terme.

Toujours dans le Tableau 7, nous indiquons en bleu un second seuil à partir duquel nous observons des SD formées par insertion d'un, deux ou trois mots au sein de l'expression recherchée. Il est donc bien possible d'observer des types de défigements différents en fonction du seuil que nous analysons. À ce stade, nous trouvons ces seuils par une lecture des résultats obtenus pour deux expressions. Un de nos prochains objectifs sera d'identifier automatiquement les seuils permettant d'isoler uniquement des SD. La finalité de ce travail nous permettra de savoir si des seuils de mesure de similarité peuvent représenter plusieurs types de défigement, comme c'est le cas avec les exemples que nous avons traités.

Potentiel défigement			Cos	Dic	Ham	Jac	Kul	Mat	Rus	#
que	la force de dieu	soit avec toi	,78	,77	,62	,62	,45	,62	,62	9
que	la force du café	soit avec toi	,78	,77	,62	,62	,45	,62	,62	6
que	la force de guérir	soit avec toi	,78	,77	,62	,62	,45	,62	,62	6
que	la force et l amour	soit avec toi	,78	,77	,62	,62	,45	,62	,62	4
que	la force du vent	soit avec toi	,78	,77	,62	,62	,45	,62	,62	3
que	la force du tigre	soit avec toi	,78	,77	,62	,62	,45	,62	,62	3
que	la force du dragon	soit avec toi	,78	,77	,62	,62	,45	,62	,62	3
que	la force	soit avec moi et toi	,78	,77	,62	,62	,45	,62	,62	2
que	la force of god	soit avec toi	,78	,77	,62	,62	,45	,62	,62	2
que	la force et la patience	soit avec toi	,76	,74	,53	,59	,42	,59	,59	4
que	là force	soit avec toi	,73	,73	,57	,57	,4	,57	,57	9
que	la paix	soit avec toi	,73	,73	,57	,57	,4	,57	,57	6
que	la force	soit en toi	,73	,73	,57	,57	,4	,57	,57	6
que	la force et le courage	soit avec toi	,73	,71	,56	,56	,38	,56	,56	6
que	le pouvoir de la force	soit avec toi	,73	,71	,56	,56	,38	,56	,56	3
que	le force	soit avec toi	,73	,73	,57	,57	,4	,57	,57	3
que	la santé	soit avec toi	,73	,73	,57	,57	,4	,57	,57	2
que	la réussite	soit avec toi	,73	,73	,57	,57	,4	,57	,57	2
que	la force	ne soit pas avec toi	,70	,69	,53	,53	,36	,53	,53	6
que	la force	ne soit jamais avec toi	,70	,69	,53	,53	,36	,53	,53	2

TABLE 7 – Résultats obtenus avec la SF « Que la force soit avec toi » pour chaque similarité, classés selon la mesure de similarité cosinus.

6 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à un cas particulier de créativité lexicale : le défigement linguistique. Nous avons construit un corpus de tweets sur lequel nous avons appliqué une méthode de reconnaissance de séquences défigées (SD) fondée sur des mesures de similarité. Nous avons exploité des mesures de similarité pour identifier des SD dans ce corpus de tweets. Nous avons utilisé différentes couches d'analyse linguistique pour parvenir à une reconnaissance de ces SD en nous limitant à une approche des défigements interprétables hors contexte. Dans le futur, nous envisageons d'exploiter plus avant la notion de « remotivation » des SD telle que proposée par (Eline & Zhu, 2014) qui pourrait permettre de différencier les SD volontaires des énoncés fautifs et d'interroger la relation entre degré de figement et défigement. Nous estimons qu'une mesure de similarité combinant les couches linguistiques brutes, étiquetée et lemmatisée permet une meilleure reconnaissance de défigements. Il nous reste à déterminer si la couche phonétique peut, elle aussi, se révéler utile dans notre tâche et à exploiter également une couche syllabique. Les résultats obtenus avec les mesures de similarité nous montrent qu'il est possible de créer une classification des défigements. On observe des seuils de similarité permettant de regrouper des SD formées par les mêmes procédés linguistiques. Nous nous demandons si ces seuils devront être « fixés » globalement ou pourront être calculés pour chaque expression, ou pour chaque procédé de construction. Déterminer automatiquement ces seuils, en exploitant des modèles de langue contextuels est une autre voie que nous comptons explorer.

Références

BAPTISTA J., CORREIA A. & FERNANDES G. (2004). Frozen Sentences of Portuguese : Formal Descriptions for NLP. In T. TANAKA, A. VILLAVICENCIO, F. BOND & A. KORHONEN, Éds., *ACL Workshop on Multiword Expressions : Integrating Processing*, p. 72–79, Barcelona, Spain. HAL : hal-01025937.

CATALÀ D. & BAPTISTA J. (2007). Spanish adverbial frozen expressions. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, p. 33–40.

CUSIMANO C. (2015). Figement de séquences défigées. *Pratiques*, (159-160), 69 78. DOI : 10.4000/pratiques.2833.

ELINE J. & ZHU L. (2014). Défigement et inférence - cas d'études du Canard enchaîné. *SHS Web of Conferences*, **8**, 681 695. DOI : 10.1051/shsconf/20140801235.

FIALA P. & HABERT B. (1989). La langue de bois en éclat : les défigements dans les titres de presse quotidienne française. *Mots. Les langages du politique*, **21**(1), 83–99. DOI : 10.3406/mots.1989.1504.

FORT K., GUILLAUME B., CONSTANT M., LEFÈBVRE N. & PILATTE Y.-A. (2018). "Fingers in the Nose" : Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 207–213, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

FORT K., GUILLAUME B., PILATTE Y.-A., CONSTANT M. & LEFÈBVRE N. (2020). Rigor Mortis : Annotating MWEs with a Gamified Platform. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4395–4401, Marseille, France : European Language Resources Association.

FRANÇOIS J. & MANGUIN J.-L. (2006). Dispute théologique, discussion oiseuse et conversation téléphonique : les collocations adjectivo-nominales au cœur du débat. *Langue française*, **150**(2), 50–65. DOI : 10.3917/lf.150.0050.

GALISSON R. (1993). Les palimpsestes verbaux : des révélateurs culturels remarquables, mais peu remarqués. *Repères. Recherches en didactique du français langue maternelle*, **8**(1), 41–62. DOI : 10.3406/reper.1993.2091.

GRECIANO (1985). Gréciano Gertrud, Signification et dénotation en Allemand. La sémantique des expressions idiomatiques, Paris, Klincksieck, 1983. *L'information grammaticale*, **24**(1), 47–48.

GROSS G. (1996). *Les expressions figées en français. Noms composés et autres locutions - Gaston Gross*. OPHRYS.

GROSS M. (1982). Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, **11**(2), 151. DOI : 10.7202/602492ar.

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

LECLER A. (2004). Blague à part, peut-on traiter la question du défigement en termes dialogiques ? *Cahiers de praxématique*, (43), 81–106. DOI : 10.4000/praxematique.1807.

LECLÈRE C. (2000). Expressions figées dans la francophonie : le projet bfqs.

MEJRI S. (1998). La conceptualisation dans les séquences figées. *L'information grammaticale*, **2**(1), 41–48. DOI : 10.3406/igram.1998.3699.

MEJRI S. (2005). Figement absolu ou relatif : la notion de degré de figement. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (53), 183–196. DOI : 10.4000/linx.283.

MEJRI S. (2009). Figement, défigement et traduction. Problématique théorique. *Pratiques*, p. 153.

MOLINARO N. & CARREIRAS M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biological Psychology*, **83**, 176–190. DOI : 10.1016/j.biopsycho.2009.12.006.

NUNBERG G., SAG I. A. & WASOW T. (1994). Idioms. *Language*, **70**(3), 491–538. DOI : 10.2307/416483.

ROMMERS J., DIJKSTRA T. & BASTIAANSEN M. (2013). Context-dependent Semantic Processing in the Human Brain : Evidence from Idiom Comprehension. *Journal of Cognitive Neuroscience*, **25**(5), 762–776. DOI : 10.1162/jocn_a_00337.

TAN M., JIANG J. & DAI B. T. (2021). A bert-based two-stage model for chinese chengyu recommendation. *Transactions on Asian and Low-Resource Language Information Processing*, **20**(6), 1–18.

VALITUTTI A., TOIVONEN H., DOUCET A. & TOIVANEN J. M. (2013). "let everything turn well in your wife" : generation of adult humor using lexical constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 243–248.