Le traitement informatique du défigement : des corpus à l'intelligence artificielle

Jean-Pierre COLSON

Résumé

Afin de mieux cerner les aspects quantitatifs du défigement, le recours au traitement informatique s'avère nécessaire. Dans cette contribution, nous nous intéressons tout d'abord aux techniques de linguistique de corpus qui peuvent être mises à profit pour le défigement. En particulier, nous montrons que le recours à de vastes corpus et à des requêtes complexes permet de mesurer en partie la proportion de défigement pour une séquence figée donnée. Les corpus ne livrent toutefois que des données partielles sur le défigement et ne permettent pas d'extraire automatiquement des séquences défigées. À cet effet, il est nécessaire de faire appel à l'intelligence artificielle et aux vastes modèles linguistiques pré-entraînés (*LLMs*). Dans une seconde expérience pratique, nous montrons comment l'usage d'un modèle *LLM* ouvre la voie à une reconnaissance et à une extraction automatique du défigement.

Mots-clés : phraséologie, figement, défigement, TAL, intelligence artificielle

Abstract

In order to better comprehend the quantitative aspects of idiom modification, the utilization of computational methods can prove highly advantageous. In this contribution, we initially delve into corpus linguistics techniques that can be harnessed for idiomatic variation analysis. Specifically, we demonstrate that employing extensive corpora and complex queries allows for the partial measurement of the proportion of variation within a given fixed sequence. However, corpora yield only partial data on idiom modification, notably falling short in automatically extracting those constructions. To address this limitation, recourse to artificial intelligence and large pre-trained linguistic models (LLMs) becomes imperative. In a subsequent practical experiment, we illustrate how the application of an LLM model paves the way for the automatic recognition and extraction of idiom modification.

Key words: phraseology, idioms, idiom modification, NLP, artificial intelligence

Introduction

Aborder le défigement d'un point de vue informatique ou automatisé n'est pas une tâche aisée. Ainsi que le montrent les diverses contributions de ce volume, le défigement implique en effet des constructions linguistiques complexes, qui relèvent de la culture, des liens avec l'actualité, de l'humour, de l'ironie, etc. D'un point de vue théorique également, Mejri (2013) a bien montré toute la complexité du phénomène, où sens figuré et sens littéral coexistent souvent.

Le traitement informatique du défigement est confronté à la grande diversité de ses réalisations concrètes. Mejri (2009) détaille par exemple les diverses modifications de la fixité qu'entraîne le défigement : fixité formelle (phonétique, morphologique, paradigmatique, syntagmatique, d'actualisation, syntaxique, de la combinatoire phrastique, de la combinatoire discursive) ; fixité sémantique (dualité littéralité / global, remotivation du sens littéral par association, remotivation par mention métalinguistique).

Le défigement représente dès lors un défi de taille pour le traitement informatique, qu'il s'agisse des corpus ou de l'intelligence artificielle. L'intelligence humaine, elle aussi, est parfois mise à rude épreuve par le défigement : la traduction dans une langue étrangère est particulièrement complexe, voire impossible (Ben Amor 2008, Mejri 2009, Pernot 2013).

Le défigement pourrait faire figure de simple curiosité linguistique : une construction plutôt rare et peu intéressante. Il n'en est rien : au contraire, les phénomènes périphériques sont souvent révélateurs de caractéristiques majeures. En linguistique, nous pourrions citer les exemples de la grammaire générative et de la grammaire des constructions, deux théories qui se sont élaborées en partie sur la base de constructions linguistiques peu courantes ou marginales.

Aujourd'hui, la linguistique est confrontée de plein fouet à la révolution de l'intelligence artificielle (IA) : la traduction automatique, autrefois de piètre qualité, s'améliore sans cesse et impressionne même les professionnels. Nombre de tâches linguistiques, liées à la sémantique ou à la syntaxe, sont à la portée de modèles informatiques complexes. Les vastes corpus linguistiques constituent depuis la dernière décennie du XX^{e} siècle une référence incontournable pour les travaux de linguistique. Or, ils sont aujourd'hui menacés par la montée en puissance de l'intelligence artificielle, car celle-ci fournit des modèles de taille plus limitée, structurés selon des algorithmes et plus riches en information (Colson 2024). La complexité linguistique du défigement permet précisément d'éclairer en partie le débat entre les corpus, le traitement automatique du langage (TAL) et l'intelligence artificielle (IA). Les modèles sont en quelque sorte poussés dans leurs derniers retranchements, car le défigement assemble souvent, de manière très subtile, les représentations morpho-phonétiques, syntaxiques, sémantiques et socioculturelles. L'IA est-elle capable de gérer tous ces éléments et les corpus linguistiques sont-ils

toujours utiles ? Les expériences sur le défigement contribuent en partie à éclairer ce débat, en raison des nombreuses facettes de cette construction linguistique. En retour, les expériences de TAL et d'IA complètent la description traditionnelle du défigement et peuvent mettre en lumière certains aspects encore peu étudiés.

Une étude pionnière s'est lancée dans la reconnaissance automatique du défigement (Bezançon & Lejeune 2023). Les auteurs y ont tenté une première expérience d'extraction automatique des séquences défigées, à partir d'un corpus de tweets. Ils se sont toutefois limités aux cas de défigement analysables hors contexte, et se sont basés sur une liste préétablie de séquences figées, à partir desquelles ils ont tenté de retrouver les cas de défigement. À cette fin, Bezançon & Lejeune ont eu recours à des mesures de similarité sémantique bien connues en TAL, telles que la Similarité Cosinus ou le score Jaccard. Leur étude fournit des résultats intéressants, mais pourrait être approfondie sur deux points. Tout d'abord, elle se limite à l'extraction de séquences défigées à partir d'une liste préétablie de séquences figées. Il ne s'agit donc pas d'une véritable extraction automatique de séquences défigées : le modèle ne permet pas de savoir, à partir d'un texte choisi au hasard, s'il contient ou non des séquences défigées, à moins que (par chance) les séquences figées correspondantes figurent dans la liste de départ. Autre réserve : les mesures de similarité sémantique utilisées dans cette étude livrent des résultats intéressants, mais la plupart des travaux d'extraction sémantique ont aujourd'hui recours à l'intelligence artificielle, bien mieux armée pour saisir toutes les subtilités des interactions sémantiques.

À partir de ces divers éléments, nous tenterons dans cette contribution d'apporter des éléments de réponse à deux questions de recherche :

1. Quelles techniques de corpus peuvent être utilisées pour l'étude du défigement ?
2. Quel est l'apport de l'intelligence artificielle à l'étude et à l'extraction des séquences défigées ?

1. Analyse du défigement dans les corpus : une méthodologie semi-automatisée

Les séquences défigées modifient une séquence figée, que l'on peut également appeler phraséologisme (Burger 2003). Il est bien connu que la fréquence des phraséologismes est très variable dans les corpus : certaines expressions idiomatiques courantes peuvent présenter un nombre relativement faible d'occurrences dans un corpus donné (Colson 2003, 2007, 2008). Ceci vaut d'autant plus pour les séquences défigées : elles apparaissent notamment dans les journaux satiriques (Sullet-Nylander 2005, Vrbinc & Vrbinc 2011, Zhu & Eline 2014) mais leur fréquence reste faible et leur forme imprévisible, car elles sont liées au contexte linguistique, socioculturel, politique etc. Autant dire que les séquences défigées sont particulièrement difficiles à détecter dans les corpus linguistiques, même si ces derniers sont de taille gigantesque.

Un point précis de l'étude du défigement peut être abordé par le biais des corpus. Selon Zhu & Eline (2014), « plus la construction d'une SF (séquence figée) est rare, plus la SF est apte à un défigement » (Zhu & Eline 2014 : 684). D'une manière plus générale, la fréquence relative des SF et les SD (séquences défigées) peut être mesurée dans les corpus, à condition qu'ils soient de taille très importante. La requête qui vise à extraire l'information linguistique du corpus doit en outre permettre une certaine souplesse. Ceci nécessite la manipulation (même rudimentaire) d'un langage informatique ou la maîtrise des outils de recherche avancée offerts par les corpus.

Qu'il s'agisse du français ou de la plupart des langues couramment utilisées dans le monde, le Sketch Engine[1] offre à la fois des corpus de taille gigantesque et des requêtes complexes. Ce merveilleux outil, bien connu des linguistes de corpus, convient donc très bien à l'étude de la phraséologie en général et peut même fournir des résultats intéressants pour les séquences défigées.

À titre d'exemples, nous avons sélectionné au hasard de nos lectures de la presse francophone cinq séquences figées qui se prêtent régulièrement à un défigement :

(1) L'insoutenable légèreté de l'être.
(2) Il / elle n'en est pas à son coup d'essai.
(3) Qui trop embrasse mal étreint.
(4) Tous les chemins mènent à Rome.
(5) Là où il y a de la gêne, il n'y a pas de plaisir.

Notre propos est seulement d'illustrer l'apport des corpus à l'étude du défigement et nous n'aborderons pas ici l'origine, les sens principaux ou le classement de ces séquences figées. Elles sont bien connues de la plupart des francophones, et pourtant seuls les plus vastes corpus du Sketch Engine livrent suffisamment de contextes et d'exemples de défigement. Nous avons eu recours à la version la plus récente, le corpus « French Web 2020 (frTenTen20), qui compte près de 18 milliards de signes (*tokens*).

Le Sketch Engine dispose de requêtes avancées en langage *CQL* (*Corpus Query Language*). Ce dernier permet de combiner des lemmes (en tenant compte des formes conjuguées, temps et déclinaisons), des intervalles entre les mots ou encore des étiquettes (*tags* : Nom, Verbe etc.). Un choix doit être opéré parmi les diverses requêtes possibles, mais il est relativement aisé d'aboutir à une requête *CQL* qui mette en lumière une séquence figée et son défigement potentiel. Ainsi, pour l'exemple (1), *l'insoutenable légèreté de l'être*, nous avons choisi comme requête l'article défini (en tant que lemme), suivi de *insoutenable*, suivi d'un substantif, de la préposition *de* et d'un autre substantif. La requête *CQL* se présente dès lors comme en (6).

[1] http://sketchengine.eu

(6)
```
[lemma="le"][lemma="insoutenable"][tag="N.*"][lemma="de"]
[lemma="le"][tag="N.*"]
```

Ceci nous livre par exemple, outre *l'insoutenable légèreté de l'être* : *l'insoutenable ambiguïté de l'aide, l'insoutenable menace de l'égalité, l'insoutenable légèreté de la folie, l'insoutenable pollution de l'air, l'insoutenable mystère de l'existence, etc.* Parmi les exemples les plus courants de défigement de cette SF, notons le jeu avec des sonorités proches de *l'être* : *l'aide, l'aile, l'air, l'art* ; ou encore la simple modification du second substantif : *l'insoutenable légèreté de la beauté, de la bourgeoisie, de la chance, de la crédibilité, de la décision, de la dette, de la fiscalité, de la globalisation, de la politique, de la téléréalité*, etc.

L'outil Sketch Engine produit l'ensemble des résultats sous la forme d'une concordance, qui peut être exportée dans un fichier csv ; celui-ci peut être lu par Excel ou équivalent. De cette manière, il est possible de mesurer la proportion de séquences défigées par rapport à la séquence figée d'origine, de manière semi-automatisée ou par le biais d'un programme informatique qui analysera le fichier csv.

Cette méthode nécessite toutefois un contrôle manuel. En effet, tous les résultats produits par la requête ne constituent pas des exemples de défigement et doivent dès lors être écartés manuellement. Voici quelques exemples fournis par la requête sous (6), qui ne constituent pas clairement un défigement de *L'insoutenable légèreté de l'être*, même si certains cas sont discutables : *les insoutenables bombardements de la côte, les insoutenables images de la maternité, L'Insoutenable Assomption de la reine bouffonne, l'insoutenable attente de la mort, l'insoutenable barbarie de l'homme*, etc.

En procédant de la sorte pour les exemples (1) à (5) mentionnés plus haut, nous pouvons présenter (Tableau 1) le nombre d'occurrences de la séquence figée (colonne frSF), de la séquence défigée (colonne frSD) et calculer ainsi le pourcentage de défigement (colonne %DF) par rapport à l'ensemble des résultats de la requête, soit frSD / (frSF + frSD).

SF	**frSF**	**frSD**	**%DF**
(1) *l'insoutenable légèreté de l'être*	745	324	30 %
(2) *il / elle n'en est pas à son coup d'essai*	3689	13	0,4 %
(3) qui trop embrasse mal étreint	540	12	2 %
(4) *tous les chemins mènent à Rome*	773	704	48 %
(5) *là où (il) y a de la gêne, (il) (n') y a pas de plaisir*	24	7	23 %

Tableau 1. Pourcentage de défigement pour 5 séquences figées (corpus frTenTen20)

Comme le montre le Tableau 1, le pourcentage de défigements pour une SF donnée varie beaucoup. La séquence figée (1) est un cliché, ce qui ouvre sans doute la

porte à bien des modifications contextuelles. Dans le cas de (2), la SF est une expression verbale, partiellement idiomatique et très courante, qui ne livre que très peu de cas de défigement. Citons, par exemple : *n'en est pas à son coup de maître, n'en est pas à sa haie d'essai, n'en est pas à son ballon d'essai*. La SF (3) peut être considérée comme un proverbe, et ne livre que 2 % de défigement, en dépit des 552 occurrences fournies par la requête. Notons toutefois que la méthodologie fondée sur les corpus permet à nouveau de trouver facilement, par le biais de la concordance et du fichier Excel, l'un ou l'autre cas manifeste de défigement : *qui trop étreint mal embrasse, qui trop impose mal étreint, qui trop palabre mal étreint, qui trop voit mal étreint*.

Le proverbe d'origine latine en (4), très courant dans le corpus, livre le pourcentage le plus élevé de défigement. Citons par exemple : *tous les chemins mènent à Amazon, tous les chemins mènent à Hollywood, tous les chemins mènent à Macron, tous les chemins mènent à Netflix, tous les chemins mènent à Ron, tous les cinémas mènent à Hollywood, tous les colimaçons mènent à Rome, tous les drones mènent à Rome, tous les débats mènent à Rome*. Le fichier de concordance fait clairement apparaître que la plupart des SD jouent sur la variation entre *Rome* et d'autres entités nommées (villes, pays, personnages ou sociétés célèbres).

En résumé, la méthodologie proposée ici pour analyser le défigement à travers les corpus est relativement simple. Tout d'abord, il s'agit de cibler une séquence figée quelconque (collocation, expression idiomatique, proverbe, cliché, etc.) et de créer une concordance à l'aide d'un outil (tel que le *Sketch Engine*) ou d'un programme informatique. La seule étape cruciale, qui exige une légère manipulation informatique, est la mise au point d'une requête de recherche, par exemple en langage *CQL* comme en (6). Ensuite, une lecture attentive et un tri de la concordance par colonne permettent aisément de déterminer, manuellement ou de manière automatique, la proportion de défigement.

Les quelques résultats proposés dans le Tableau 1 avaient pour seul objectif d'illustrer la méthodologie, et leur examen plus approfondi sortirait du cadre de cette contribution. Ils suffisent néanmoins à indiquer que le taux de défigement par rapport à la forme canonique de la séquence figée est très variable. Ceci tient à des facteurs complexes que d'autres travaux basés sur de vastes corpus pourraient permettre de préciser.

Nous pouvons ainsi répondre à la première question de recherche formulée dans l'introduction : les techniques de corpus utiles à l'étude du défigement combinent le travail automatisé et le travail manuel. Elles mettent en lumière la grande richesse des exemples et donnent une idée générale de la fréquence du phénomène, mais se heurtent également à certaines limites. La méthodologie de corpus proposée plus haut ne permet pas, en effet, d'extraire automatiquement les séquences défigées à partir d'un texte donné. Un tel objectif nécessite le recours à l'intelligence artificielle, comme nous tentons de l'illustrer au point suivant.

2. Intelligence artificielle et défigement : pistes de recherche

Dans cette même revue, Mayaffre & Vanni (2022) ont déjà proposé une introduction abordable à l'intelligence artificielle et au *deep learning* (apprentissage profond) dans le cadre de la linguistique textuelle. Nous renvoyons à cette étude et à Mayaffre & Vanni (2021) pour un aperçu de l'IA et de ses divers apports à l'étude des textes.

Nous nous garderons ici d'approfondir les aspects techniques et présenterons deux expériences à la portée des chercheurs qui maîtrisent les bases de l'IA appliquée à la linguistique. Parmi les divers modèles utilisés en IA, les *LLMs* (*Large Language Models*, Grands modèles linguistiques), basés sur l'architecture Transformer connaissent aujourd'hui un énorme succès.

L'architecture *Transformer* (Vaswani et al. 2017) est aujourd'hui utilisée par la plupart des chercheurs en linguistique computationnelle, en traduction automatique et même par chatGPT[2] (dont la lettre « t » est précisément l'abréviation de « Transformer » : *Generative Pretrained Transformer*).

La véritable révolution engendrée par l'architecture *Transformer* réside dans une meilleure capacité à tenir compte des contextes précis de chaque mot. Une nouvelle implémentation informatique permet au modèle de s'entraîner et de s'améliorer progressivement en tenant compte de larges contextes (des phrases entières). De cette manière, le poids relatif de chaque mot dans une phrase est pris en compte, ce qui engendre une représentation sémantique très fine.

L'architecture *Transformer* considère que le sens et les associations sont liés au large contexte et aux récurrences entre les mots. Elle est dès lors tout à fait compatible avec la sémantique distributionnelle (Baroni 2013, Boleda 2020, Emerson 2020, Erk 2012, Fabre 2015, Heylen et Bertels 2016, Lenci 2018, Mitchell & Lapata 2010), la phraséologie (Burger 2003) et le figement lexical (Mejri 2003), même si aucune de ces références n'est citée dans l'article fondateur (Vaswani et al. 2017).

En résumé, les *LLMs*, vastes modèles linguistiques basés sur *Transformer*, sont très bien armés pour gérer l'idiomaticité. Les expériences récentes l'ont d'ailleurs démontré. Ainsi, lors de l'édition 2020 de la tâche partagée Parseme 1.2., consacrée à l'extraction des expressions verbales (Ramisch et al. 2020), les meilleurs résultats obtenus étaient basés sur les *LLMs*. Ces derniers permettent en effet le *fine-tuning* ou affinage du modèle.

En termes simples, les *LLMs* constituent, pour une langue, une représentation sémantique et combinatoire de l'ensemble des mots du corpus. Dans le modèle, les

[2] http://www.openai.com

principaux contextes et phrases d'utilisation de chaque mot sont convertis en une sorte de signature numérique (quelques centaines de nombres réels pour chaque mot). Dès lors, le modèle offre une représentation cohérente de l'usage et de la signification des mots de la langue. Il peut être mis à profit pour accomplir diverses tâches linguistiques, telles que la traduction, l'extraction de liens sémantiques ou d'expressions idiomatiques.

À ces fins, le modèle peut être ajusté, *affiné* en lui livrant simplement un nouveau jeu de données d'apprentissage. Un minimum de connaissances en informatique (des bases de programmation, surtout en langage Python) suffit à réaliser un tel affinage d'un *LLM* : les programmes simples (scripts) sont accessibles sur le Web, et l'entraînement du modèle peut se faire via des plateformes gratuites[3].

Nous avons proposé (Colson 2024) une telle méthodologie pour l'extraction de trois grandes catégories phraséologiques (expression idiomatique, collocation, formule communicative). Les résultats prometteurs suggèrent que l'affinage d'un *LLM* permet effectivement à l'intelligence artificielle de reconnaître ces constructions idiomatiques à partir d'un texte quelconque.

Afin de répondre à notre seconde question de recherche posée dans l'introduction (l'apport de l'intelligence artificielle à l'étude et à l'extraction des séquences défigées), nous présentons une expérience simple d'apprentissage du défigement par un modèle *Transformer* affiné. Notre objectif est de mesurer la capacité d'un tel modèle à reconnaître le défigement, à partir d'un jeu de données limité, qui puisse être complété par des données plus importantes.

Nous avons constitué un jeu de données d'apprentissage composé de 400 exemples pris au hasard dans la presse francophone, en veillant à une représentation équilibrée de figements et de défigements. Concrètement, trois catégories étaient présentes dans les données d'apprentissage : FIGEMENT (exemple : *un tramway nommé désir*), DÉFIGEMENT (exemple : *un acteur nommé désir*) et AUTRE (exemple : *un tramway bloquait le passage*). Comme ces trois exemples l'indiquent, les tests préliminaires ont mis en lumière qu'il était indispensable d'entraîner le modèle d'IA en lui donnant des données qui présentaient des éléments proches mais clairement différenciées par le figement, le défigement ou l'absence des deux (catégorie « AUTRE »). Alors que le modèle peinait à reconnaître les cas de défigement, l'ajout de séquences de ce type s'est révélé crucial. D'autres exemples du même type, issus des données d'apprentissage du modèle, sont présentés dans le Tableau 2.

Exemples	**catégorie**
la victoire en chantant	FIGEMENT

[3] Par exemple : Google Colab (https://colab.research.google.com) ou Kaggle (https://www.kaggle.com/)

la victoire en déchantant	DÉFIGEMENT
la victoire en peu de temps	AUTRE
un tramway nommé désir	FIGEMENT
un acteur nommé désir	DÉFIGEMENT
un tramway bloquait le passage	AUTRE
une étoile est née	FIGEMENT
une étoile aînée	DÉFIGEMENT
une étoile brillait dans	AUTRE
il n'en est pas à son coup d'essai	FIGEMENT
il n'en est pas à son coup d'excès	DÉFIGEMENT
il n'en est pas tout à fait conscient	AUTRE
le cyclisme sur route	FIGEMENT
le cyclisme sur doutes	DÉFIGEMENT
le cyclisme et le football	AUTRE
qui voyage loin ménage sa monture	FIGEMENT
qui voyage loin ménage sa voiture	DÉFIGEMENT
qui voyage loin doit prendre des vivres	AUTRE

Tableau 2. Exemples de données d'apprentissage du modèle d'IA pour le défigement

Comme illustré par les exemples du Tableau 2, la mise au point du modèle d'apprentissage a nécessité un travail manuel de sélection. À chaque fois, la séquence défigée a été introduite dans le modèle, accompagnée de la séquence figée correspondante et d'une séquence libre, relativement proche des deux autres.

L'entraînement du modèle par affinage d'un modèle existant, à l'aide des 400 exemples d'apprentissage, est relativement simple d'un point de vue technique[4]. Le Tableau 3 présente les résultats obtenus par catégorie (FIGEMENT, DÉFIGEMENT, AUTRE) et en moyenne.

	précision	**rappel**	**score f1**
FIGEMENT	0,75	1,00	0,86
DÉFIGEMENT	1,00	0,54	0,70
AUTRE	0,83	0,71	0,77

[4] Pour l'entraînement du modèle, nous avons pu utiliser, via la plateforme Google Colab (https://colab.research.google.com) un GPU de 90 GB, sur une machine distante Tesla T4. Nous avons eu recours à Tensorflow et Keras. Le modèle pré-entraîné était "dbmdz/electra-base-french-europeana-cased-discriminator", disponible sur Huggingface (hugginface.co). L'entraînement a été réalisé sur 15 époques, à un taux de 2^{e}-5. Le script en Python a utilisé la librairie Transformer, dont les fonctions AutoModel, AutoTokenizer.

Moyenne pondérée[5]	0,84	0,80	0,79

Tableau 3. Résultats obtenus par le modèle d'IA pour le défigement

Le tableau 3 rapporte, dans les catégories utilisées lors de l'expérience (FIGEMENT, DÉFIGEMENT, AUTRE) la précision, le rappel et le score f1. Pour l'exprimer de manière simple, la précision répond à la question « le modèle dit-il à chaque fois la vérité ? », alors que le rappel correspond à la question « le modèle dit-il toute la vérité ? ». Par exemple, pour la catégorie « AUTRE », une précision de 0,83 indique que dans 83 % des cas, le modèle avait raison lorsqu'il classait une séquence dans cette catégorie. Par contre, si l'on tient compte de toutes les séquences que le modèle aurait dû classer dans cette catégorie, le score de rappel de 0,71 nous indique que, parmi toutes les séquences qui auraient dû recevoir l'étiquette « AUTRE », seules 71 % l'ont effectivement reçue. Le score f1, quant à lui est la moyenne harmonique[6] entre la précision et le rappel.

Comme le Tableau 3 le montre, notre expérience d'affinage d'un modèle *Transformer* afin de reconnaître les séquences défigées obtient une moyenne globale pondérée tout à fait acceptable : 84 % de précision et 80 % de rappel. Le modèle a pu, à chaque fois, affirmer à juste titre qu'une séquence était défigée (100 % de précision) mais n'a pas pu identifier toutes les séquences défigées qu'il devait reconnaître : le rappel est de 54 % pour les séquences défigées.

Cette expérience ne portait que sur 400 exemples et ses résultats doivent dès lors être interprétés avec prudence : d'autres expériences, basées sur la collecte de milliers de séquences, pourraient préciser les résultats. Notre objectif était également limité : donner un premier aperçu de l'apport possible de l'IA à l'étude du défigement. De ce point de vue, les résultats sont prometteurs : le recours à l'IA et au *deep learning* livre manifestement des résultats intéressants. Le modèle de taille modeste testé ici obtient déjà un score f1 de 86 % pour la reconnaissance des séquences figées et de 70 % pour les séquences défigées.

Le Tableau 4 présente, à titre d'exemples, quelques séquences prises au hasard de nos lectures après l'entraînement du modèle. Elles ne figuraient donc pas dans les données d'entraînement et ont pu être correctement identifiées par le modèle.

Exemples	**catégorie**
il se Taipei notre tête	DÉFIGEMENT
il se paie notre tête	FIGEMENT
dans quel monde vit-on ?	FIGEMENT

[5] La moyenne pondérée tient compte du nombre d'exemples dans chaque catégorie.

[6] La moyenne harmonique est l'inverse de la moyenne arithmétique des inverses des termes.

mais dans quel monde Vuitton ?	DÉFIGEMENT
la fosse aux serpents	FIGEMENT
la fosse aux serments	DÉFIGEMENT
la fosse aux animaux	AUTRE
mariage de raison	FIGEMENT
ratages de raison	DÉFIGEMENT

Tableau 4. Exemples de séquences correctement identifiées par le modèle d'IA

Nous pouvons retenir de cette expérience d'affinage d'un vaste modèle linguistique deux points principaux. Tout d'abord, l'intelligence artificielle permet, même avec un nombre relativement limité d'exemples, d'aboutir assez rapidement à un classement acceptable de séquences figées, défigées et autres. Nos résultats suggèrent que le modèle pourrait être amélioré par des données d'apprentissage plus nombreuses. Dans le cas du défigement, il est toutefois difficile d'aboutir à une collection de milliers d'exemples. Les modèles plus limités tel que celui présenté plus haut pourraient permettre de constituer progressivement une base de données du défigement, en traitant les textes de manière automatisée.

Un autre point important ressort de notre expérience d'IA : le modèle permet de reconnaître le défigement, mais seulement au prix d'exemples contrastés qui jouent sur tous les aspects formels du défigement. Seuls ces derniers sont identifiés aisément par les vastes modèles linguistiques, qui ne peuvent tenir parfaitement compte de toutes les nuances culturelles, sémantiques ou liées à l'actualité.

Notre expérience de reconnaissance du défigement par l'IA suggère toutefois que l'idiomaticité en général constitue bel et bien l'un des points forts de l'architecture *Transformer*, à condition de pouvoir l'utiliser de manière cohérente.

Afin d'illustrer ce dernier élément, il peut être utile de nous tourner vers DeepL[7] qui, rappelons-le, est également basé sur l'IA et *Transformer*. La traduction automatique neuronale peut-elle fonctionner à partir de séquences défigées du français ?

Ce domaine de recherche mérite également d'être étudié à plus grande échelle, car le défigement constitue un problème majeur de compréhension linguistique et de traduction. À son tour, le défigement pourrait contribuer à mieux analyser les limites de l'IA et éventuellement de l'améliorer en matière de traduction automatique.

Dans l'attente de données plus nombreuses à ce sujet, nous avons réalisé une autre expérience limitée, à partir de cinq exemples de défigement formel

[7] https://deepl.com

récemment utilisés dans l'actualité. Nous précisons à chaque fois la source et la signification des séquences défigées.

(7) Cabinets de conseils : le gouvernement n'en est pas à son coup d'excès. (*Libération*, 11/07/2023).
Explication : jeu de mots avec *coup d'essai*, qui fait partie de la séquence figée *il n'en est pas à son coup d'essai*. En remplaçant *coup d'essai* par *coup d'excès*, le journal adopte un point de vue critique par rapport au recours du gouvernement français à des cabinets de conseil.

(8) Présidentielle en Équateur : pour la gauche, le retour est joué ? (*Libération*, 20/07/2023).
Explication : jeu de mots entre *retour* et *tour*, qui fait partie de la séquence figée *le tour est joué*. En remplaçant *tour* par *retour*, le journal adopte un titre plus original, qui signale que la gauche reviendrait au pouvoir.

(9) (À propos du président E. Macron) Comme s'il s'adressait aux Français en leur disant « Baignez-vous, il n'y a rien à voir… ». (*Le Soir*, 21/07/2023)
Explication : utilisation de *Baignez-vous* au lieu de *Circulez*, qui fait partie de la séquence figée *Circulez, il n'y a rien à voir*. En remplaçant *Circulez* par *Baignez-vous*, le journal fait allusion à la période estivale.

(10) Le fond de l'air effraie (*Libération*, 05/12/2023).
Explication : jeu de mots entre *effraie* et *est frais*, qui fait partie de la séquence figée *le fond de l'air est frais*. Le journal fait ici allusion à la montée de l'extrême-droite en France : l'expression *le fond de l'air est frais* est employée au sens figuré, et le sens exact est précisé par la substitution *est frais / effraie*.

(11) Depardieu : monstre saqué (*Le Soir*, 27/11/2023).
Explication : jeu de mots entre *saqué* et *sacré*, qui fait partie de la séquence figée *monstre sacré*. Le journal fait ici allusion à la campagne de presse qui a dénoncé les agissements de l'acteur français Gérard Depardieu.

Les exemples 7 à 11, caractérisés par des défigements formels, posent de sérieux problèmes à la traduction automatique neuronale. Ainsi, DeepL[8] livre des traductions anglaises acceptables (avec perte du jeu de mots) pour (7) et (8) mais erronées pour (9), (10) et (11). Faute d'espace, nous ne nous attarderons pas aux détails de ces traductions, qui figurent en note 8. La perte du jeu de mots est le principal écueil lors de la traduction du défigement ; ainsi, dans l'exemple (11), le *monstre saqué* devient en anglais *sapped monster* (monstre épuisé), car DeepL ne

[8] deepl.com, consulté la dernière fois le 16/01/2024. Les traductions en anglais étaient : (7) Consultancy firms: this is not the first time the government has gone overboard. (8) Presidential election in Ecuador: for the left, it's back to the drawing board? (9) It's as if he were saying to the French: "Swim, there's nothing to see...". (10) The air is frightening. (11) Depardieu: a sapped monster.

reconnaît pas l'expression *monstre sacré* (qui existe pourtant en anglais aussi : *sacred monster*).

Comme cette brève expérience le montre, le jeu linguistique subtil entre figement et défigement continue de poser problème à la traduction automatique. Les études quantitatives du défigement devraient permettre de contribuer à une meilleure gestion de l'idiomaticité par la traduction neuronale.

3. Conclusions

L'étude du phénomène du défigement est complexe, et le recours à des données quantitatives permet de compléter l'ensemble du tableau. La linguistique de corpus tout d'abord, offre plusieurs méthodes manuelles ou semi-automatiques aux chercheurs en quête de données de fréquence sur le défigement. Dans une première expérience, nous avons montré que les concordanciers accessibles via le Sketch Engine, en particulier, ainsi que les requêtes complexes qu'offre cette plateforme, fournissent de précieux renseignements sur la fréquence des défigements ou la proportion entre séquences figées et défigées.

La linguistique de corpus subit aujourd'hui de plein fouet la concurrence de l'intelligence artificielle, qui parvient souvent à obtenir de meilleurs résultats à partir de corpus de taille plus limitée. Dans une seconde expérience, nous avons montré que les *LLMs* et l'intelligence artificielle permettent de reconnaître en grande partie les cas de défigement. En complément des données issues de corpus, ceci devrait permettre à l'avenir de mieux cerner tous les aspects quantitatifs du défigement et d'améliorer certains aspects de la traduction automatique.

Jean-Pierre COLSON (Université de Louvain-la-Neuve)

Bibliographie

BARONI, Marco (2013), « Composition in distributional semantics », *Language and Linguistics Compass* 7, pp. 511-522.

BEN AMOR, Thouraya (2008), « Défigement et traduction intralinguale et interlinguale », *Méta* , 53 (2), pp. 443-455.

BEZANÇON, Julien & LEJEUNE, Gaël (2023). « Reconnaissance de défigements dans des tweets en français par des mesures de similarité sur des alignements textuels ». 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, 2023, Paris, France. pp.56-67. HAL Id: hal-04130174.

BOLEDA, Gemma (2020), « Distributional Semantics and Linguistic Theory », *Annual Review of Linguistics*, 6, pp. 213-234.

BURGER, Harald (2003), *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin : Erich Schmidt Verlag.

COLSON, Jean-Pierre (2003), « Corpus Linguistics and Phraseological Statistics: a Few Hypotheses and Examples », *in* Burger, Harald, Häcki Buhofer, Annelies & Gertrud Gréciano (dirs.), *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zu Methodologie und Kulturspezifik der Phraseologie*. Baltmannsweiler: Schneider Verlag Hohengehren, pp. 47-59.

COLSON, Jean-Pierre (2007), « The World Wide Web as a corpus for set phrases », *in* Burger, Harald, Dobrovol'skij, Dmitrij, Kühn, Peter & Neal R. Norrick (dirs.), *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research. Volume 2*. Berlin / New York : Walter de Gruyter, pp. 1071-1077.

COLSON, Jean-Pierre (2008), « Cross-linguistic phraseological studies: An overview », *in* Granger, Sylviane & Fanny Meunier (dirs.), *Phraseology. An interdisciplinary perspective*, Amsterdam / Philadelphia : John Benjamins, pp. 191-206.

COLSON, Jean-Pierre (2016), « Set phrases around globalization : an experiment in corpus-based computational phraseology », *in* Alonso Almeida, Francisco, Ortega Barrera, Ivalla, Quintana Toledo, Elena & Margarita E. Sánchez Cuervo (dirs.), *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics*, Newcastle: Cambridge Scholars Publishing, pp. 141-152.

COLSON, Jean-Pierre (2024), (à paraître) « Multi-word Units in Neural Machine Translation: why the Tip of the Iceberg Remains Problematic ». *In*: J. Monti, G. Corpas Pastor, R. Mitkov & C. Hidalgo Ternero (Eds.), *Recent Advances in Multiword Units in Machine Translation and Translation Technology*. Amsterdam / Philadelphia: John Benjamins.

ERK, Katrin (2012), « Vector Space Models of Word Meaning and Phrase Meaning: A Survey », *Linguistics and Language Compass*, 6, pp. 635-653.

EMERSON, Guy (2020), « What are the Goals of Distributional Semantics? », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Edition en ligne, pp. 7436-7453.

FABRE, Cécile (2015), « Sémantique distributionnelle automatique : la proximité distributionnelle comme mode d'accès au sens », *Études de linguistique appliquée*, 180, pp. 395-405.

HEYLEN, Kris & BERTELS, Ann (2016), « Sémantique distributionnelle en linguistique de corpus », *Langages*, 201, pp. 51-64.

LENCI, Alessandro (2018), « Distributional models of word meaning », *Annual review of Linguistics*, 4, pp. 151-171.

MARKANTONATOU, Stella, McCRAE, John, MITROVIĆ, Jelenan, TIBERIUS, Carole, RAMISCH, Carlos, VAIDYA, Ashwini, OSENOVA, Petya & SAVARY, Agata (éds.) (2020), « Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions », *Proceedings of the*

Joint Workshop on Multiword Expressions and Electronic Lexicons, Association for Computational Linguistics, pp. 107-118, https://aclanthology.org/2020.mwe-1.14.

MAYAFFRE, Damon & VANNI, Laurent (éds.) (2021), *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Paris, Champion.

MAYAFFRE, Damon & VANNI, Laurent (2022), « Du texte profond. Textualité et deep learning », *Le Français moderne*, pp. 135-153.

MEJRI, Salah (1997), *Le figement lexical : Descriptions linguistiques et structuration sémantique*. Thèse de doctorat. Tunis : Publications de la Faculté des Lettres de la Manouba.

MEJRI, Salah (2003), « Le figement lexical », *Cahiers de Lexicologie* 82, pp. 23-39.

MEJRI, Salah (2006), « Polylexicalité, monolexicalité et double articulation : la problématique du mot », *Cahiers de Lexicologie* 89, pp. 209-221.

MEJRI, Salah (2008), « La traduction des jeux de mots », *Equivalences* 35, Bruxelles, Editions Hazard, pp. 71-84.

MEJRI, Salah (2009), « Figement, défigement et traduction. Problématique théorique », *in* Mejri, Salah & Pedro Mogorrón Huerta (réd.), *Fijación, desautomatización y traducción / Figement, défigement et traduction*, Alicante, Université d'Alicante, pp. 153-163.

MEJRI, Salah (2013), « Figement et défigement : problématique théorique », *Pratiques* [En ligne], pp. 159-160 | 2013. Url : http://journals.openedition.org/pratiques/2847.

MITCHELL, Jeff & LAPATA, Mirella (2010), « Composition in distributional models of semantics », *Cognitive science*, 34, pp. 1388-429.

PERNOT, Caroline (2013), « Le défigement de phrasèmes pragmatiques et sa traduction », *Pratiques* [En ligne], pp. 159-160.

SULLET-NYLANDER, Françoise (2005), « Jeux de mots et défigements à La Une de Libération », *Langage et Société* 112, pp. 111-139.

VASWANI, Ashish, SHAZEER, Noam., PARMAR, Niki, USZKOREIT, Jakob, JONES, Llion, GOMEZ, Aidan N., KAISER, Łukasz & POLOSUKHIN, Illia (2017), « Attention Is All You Need ». Online edition, https://doi.org/10.48550/arXiv.1706.03762.

VRBINC, Alenka & VRBINC, Marjeta (2011), « Creative use of idioms in satirical magazines », *Jezikoslovlje* 12, pp. 75-91.