Verbum, 45(1), 2023, p. 77-98

Sens phraséologique ou distributionnel

Démêler l'écheveau sémantique par l'algorithmie

Jean-Pierre Colson
Université catholique de Louvain (Belgique)

Résumé

Établir une distinction entre associations phraséologiques d'une part et associations sémantiques de l'autre s'avère particulièrement complexe. Dans cette contribution, nous présentons les résultats de quelques expériences que nous avons menées, avec pour objectif de jeter un éclairage nouveau sur cette interaction subtile, par le biais de techniques récentes héritées du TAL. Les résultats le montrent : en dépit de plusieurs clivages théoriques entre phraséologie et sémantique distributionnelle, l'application d'algorithmes similaires peut contribuer à une fertilisation croisée entre les deux disciplines.

Abstract

Making the difference between phraseological associations on the one hand, and semantic associations on the other is particularly complex. In this paper, we report the results of a few experiments that we have carried out in order to shed fresh light on this complex interplay by means of recent NLP techniques. The results suggest that, in spite of a number of theoretical differences between phraseology and distributional semantics, the application of similar algorithms may result in cross-fertilization between the two disciplines.

1. Introduction

Charles Bally le notait déjà dans son *Traité de Stylistique française* : « la phraséologie est, elle aussi, en conflit perpétuel avec la synonymie » (1921 : 144). Il étayait notamment son propos par la paire de mots *merveille* et *miracle* : leur sens diffère mais peut se rejoindre dans des constructions phraséologiques telles que *faire merveille*. De même, un mot peut être vieilli (par exemple *nues* au sens de *nuages*), mais faire partie d'une locution familière : *tomber des nues* (Bally 1921 : 146).

Cette brillante intuition du père de la phraséologie française est particulièrement à l'ordre du jour. Les recherches de pointe en TAL (traitement automatique du langage) sont en effet confrontées à la complexité des réseaux sémantiques, dans lesquels la phraséologie interfère également.

La grande majorité des recherches sémantiques en TAL s'inscrivent dans la perspective distributionnelle. Rappelons que, même dans ses versions les plus récentes (Emerson 2020), la sémantique distributionnelle se fonde sur un principe fondamental : le sens de toute construction linguistique dépend de sa distribution, c'est-à-dire des contextes dans lesquelles elle est utilisée. Ainsi, les mots similaires apparaîtront dans des contextes similaires. Comme le soulignent de nombreuses études, principalement en langue anglaise (Baroni 2013, Boleda 2020, Emerson 2020, Erk 2012, Fabre 2015, Heylen & Bertels 2016, Lenci 2018, Mitchell & Lapata 2010), deux courants sont considérés comme les précurseurs de la sémantique distributionnelle actuelle : le structuralisme américain (Harris 1954) et la lexicologie britannique (Firth 1957).

Selon l'hypothèse distributionnelle de Harris (1954), une similarité de sens entraîne une similarité dans la distribution linguistique. Ainsi, des mots ou expressions proches d'un point de vue sémantique (par exemple, *correct* et *exact*) se retrouveront majoritairement dans des contextes similaires. La sémantique distributionnelle cite souvent (Emerson 2020) la célèbre phrase de Firth (1957 : 11), "*You shall know a word by the company it keeps*". (L'on reconnaîtra un mot par ses fréquentations, notre traduction).

Il existe un lien sémantique clair, celui de la synonymie partielle, entre *exact* et *correct* ; de même, *animal* présente un lien d'hyperonymie avec *chat* ou *chien*. Par contre, en matière de phraséologie, les associations sémantiques sont beaucoup moins évidentes : dans le dicton *l'espoir fait vivre*, l'on peut postuler un lien sémantique entre l'espoir et la vie. Si nous considérons par contre l'expression familière *à fond la caisse* (à toute vitesse), très empreinte de phraséologie, l'association sémantique est beaucoup plus complexe, car elle vaut entre *à fond* (à toute allure) et *caisse* au sens figuré et familier de voiture ; par contre, *fond* et *caisse*, pris isolément, ne sont pas associés sémantiquement et les contextes de ces deux mots ont peu de chances d'être partagés, en dehors des exemples phraséologiques de l'expression *à fond la caisse*. D'une manière

générale, les associations phraséologiques entre les mots sont imprévisibles et ne présentent pas de lien sémantique clair. Pour la plupart des expressions idiomatiques, les hasards de l'étymologie, de l'histoire et des traditions nous livrent bien des incohérences sémantiques. Ainsi, *vendre* et *mèche* ne présentent aucun lien sémantique évident, mais ils se retrouvent dans l'expression courante *vendre la mèche* (trahir un secret), par déformation d'*éventer la mèche* (découvrir la mèche d'une mine ennemie et donc en révéler la présence). En résumé, *vendre* et *mèche* ne sont en rien liés sémantiquement.

Le sens est donc tantôt distributionnel (par similarité sémantique), tantôt phraséologique (par pure convention et figement) mais parfois aussi les deux simultanément. Lorsque la sémantique distributionnelle invoque les travaux du linguiste britannique John Rupert Firth, elle crée aussi une certaine ambiguïté. Une lecture attentive de ses travaux ne laisse en effet planer aucun doute sur le sens de la phrase *You shall know a word by the company it keeps* (1957 : 11). L'auteur y vise bel et bien les associations phraséologiques (les collocations) et non le sens distributionnel. Firth illustre en effet cette affirmation par le mot anglais *ass* (au sens d'un âne) et fait précisément remarquer que ce mot se trouve, pour le citer, en compagnie familière, dans les expressions *you silly ass* (qui pourrait se traduire : espèce d'idiot !), *he is a silly ass* (c'est un âne, un imbécile) et *don't be such an ass* (ne sois pas ridicule). Nous nous situons donc tout à fait dans le domaine phraséologique. Comme pour dissiper encore toute ambiguïté, Firth précise d'ailleurs à la même page : "*From the preceding remarks, it will be seen that collocation is not to be interpreted as 'context', by which the whole conceptual meaning is implied.*" (Comme le montrent les remarques qui précèdent, les collocations ne doivent pas être identifiées au *contexte*, qui recouvre l'ensemble de la signification conceptuelle, notre traduction).

Pour Firth, les collocations sont bel et bien indépendantes du contexte, qu'il définit comme le domaine de la sémantique (« la signification conceptuelle »). En termes plus modernes, nous dirions que Firth vise dans son adage *You shall know a word by the comparny it keeps* le hasard des affinités entre les mots : *rire jaune, une critique acerbe, lâcher la proie pour l'ombre*. Celles-ci ne reposent pas sur des associations conceptuelles et extralinguistiques (comme entre *parents* et *enfants*). Il est dès lors paradoxal de voir Firth invoqué par la sémantique distributionnelle, qui ne s'attache pas aux associations phraséologiques mais bien aux associations conceptuelles. Cette matière reste toutefois complexe, car les locuteurs natifs eux-mêmes, lorsqu'ils doivent associer spontanément des mots lors d'expériences psycholinguistiques, confondent allègrement les liens sémantiques et phraséologiques. Nous prendrons pour exemple l'une des bases de données les plus fiables en la matière, réalisée pour la langue anglaise. Il s'agit des *Free Association Norms* (normes d'association libres) réalisées par l'Université de Floride du Sud (Nelson *et al.* 1998) et considérées comme particulièrement fiables pour les études psycholinguistiques. Or, pour la plupart des mots repris

dans la base de données, la phraséologie se mêle aux associations conceptuelles ou culturelles. Ainsi, *French* est associé à *Spanish, Latin, France* par les locuteurs natifs, mais aussi à *fried* (qui évoque la collocation *fried potatoes* ou *French fries*, les frites) ou encore à *kiss*, où l'on peut se demander quelle est la part de l'expression idiomatique *French kiss*.

Ces divers exemples le montrent : il est difficile de démêler l'écheveau sémantique, entre réseaux conceptuels, culturels et idiomatiques. Il est pourtant utile de faire la part des choses, notamment dans le domaine de la traduction, où tant la machine que le traducteur humain doit comprendre toutes les nuances sémantiques et culturelles du texte source et les restituer en langue cible.

L'algorithmie est au centre des recherches les plus récentes en sémantique distributionnelle, comme nous l'avons souligné plus haut. Ceci vaut également pour la phraséologie computationnelle (Corpas Pastor & Colson 2020).

Nous pouvons dégager de cette brève introduction une hypothèse de travail : le sens phraséologique et le sens distributionnel présentent de nombreux points d'interaction. Dans cette contribution, nous tenterons de vérifier cette hypothèse par des expériences récentes que nous avons menées en phraséologie computationnelle et en sémantique distributionnelle.

2. Algorithmie et extraction phraséologique : Parseme 2020

Le projet international Parseme[1] se consacre à l'extraction automatique des « expressions multi-mots », une notion largement équivalente à celle d'unités phraséologiques (Mejri *et al.* 2020). Dans ce cadre, des tâches partagées (*shared tasks*) sont régulièrement organisées lors des grands colloques internationaux de linguistique computationnelle. Elles sont l'occasion de proposer de nouveaux algorithmes d'extraction phraséologique, dont l'efficacité est mesurée objectivement à partir de données de référence validées par des locuteurs natifs.

L'édition 2020 de la tâche partagée Parseme (édition 1.2.)[2] était organisée à l'occasion du colloque international Coling 2020. Comme les éditions précédentes (1.0 et 1.1), elle était consacrée à l'extraction des expressions verbales « multi-mots ». Les 14 langues retenues pour l'édition 2020 étaient les suivantes : allemand, basque, chinois, français, grec, hébreu, hindi, irlandais, italien, polonais, portugais, roumain, suédois et turc. Notons l'absence de l'anglais.

1. https://typo.uni-konstanz.de/parseme/
2. http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_02_MWE-LEX_2020___lb__COLING__rb__&subpage=CONF_40_Shared_Task

Nous avons présenté (Colson 2020a), à l'occasion de Parseme 2020, un nouveau modèle pour l'extraction automatique de la phraséologie, *HMSid* (*Hybrid Multi-layer System for the extraction of Idioms*). Le modèle se fonde sur un score statistique que nous avons proposé, le *score-cpr* (Colson 2017, 2018a, 2018b). Ce score, qui varie de 0 à 1, repose sur la distance moyenne qui sépare les mots (*tokens*) d'une unité phraséologique polylexicale. Une division est effectuée entre le nombre d'occurrences avec une fenêtre large (entre 20 et 50 mots, selon la langue et le corpus) et une fenêtre étroite (1 à 3 mots). Prenons un exemple simple. Si, dans un corpus donné, nous trouvons 80 occurrences de l'expression *cracher le morceau* avec une fenêtre étroite et 100 occurrences avec une fenêtre large, le *score-cpr* sera de 0,80. Un système d'indexation optimisée du corpus permet de calculer très rapidement le score *cpr* (Colson 2017, 2018). L'objectif de ce score est de simuler par l'algorithmie le principe d'attraction phraséologique : un locuteur natif associera *cracher* à *morceau* en fonction de sa maîtrise de la phraséologie. De même, un algorithme peut simuler cette attraction par des méthodes statistiques.

Notre modèle *HMSid* proposé pour la langue française lors de Parseme 2020 s'est appuyé autant que possible sur les principes théoriques de la phraséologie afin d'extraire automatiquement les expressions verbales à partir du jeu de données. En l'occurrence, il s'agissait de repérer et de classer les expressions verbales selon l'attraction phraséologique, mesurée par le *score-cpr,* dans le contexte des verbes du corpus étiqueté. Nous n'abordons pas ici en détails les aspects techniques de cette expérience (voir à ce sujet Colson 2020a), mais présentons les principaux résultats et les conclusions que nous pouvons en tirer.

Nous avons tout d'abord utilisé comme corpus de référence la version française du corpus WaCky[3] (Baroni *et al.* 2009). Ce corpus ne faisait pas partie des données d'apprentissage fournies par la tâche partagée ; dès lors, les résultats fournis par ce modèle sont classés dans la « section ouverte » (*open track*). Pour la section dite fermée (*closed track*), seules les données d'apprentissage fournies par la tâche partagée peuvent être utilisées.

Le Tableau 1 ci-dessous présente les résultats globaux obtenus par notre modèle *HMSid* lors de Parseme 2020.

Tableau 1. – Scores obtenus par le modèle *HMSid* (Parseme 2020)

EMM non vues				Score global EMM				Score global (tokens)			
P	R	F1	Rang	P	R	F1	Rang	P	R	F1	Rang
27,73	53,33	**36,49**	4	63,85	67,84	**65,79**	5	66,4	67,81	**67,1**	5

3. Il s'agit du corpus « frTenTen17 » accessible via le Sketch Engine (sketchengine.eu). Ce corpus de 5,752 milliards de mots (tokens) a été assemblé par les auteurs à partir du Web, selon une méthodologie stricte inspirée de la linguistique de corpus.

Le Tableau 1 présente la synthèse des résultats globaux du modèle *HMSid* lors de la tâche partagée Parseme 2020. Pour les EMM (expressions multi-mots), un score global est calculé au centre du tableau pour l'adéquation exacte de l'extraction automatisée (par exemple l'extraction des trois mots pour *cracher le morceau*) ; le score de la partie droite (*tokens*) autorise un élément manquant dans l'extraction, par exemple la reconnaissance d'une association entre *cracher* et *morceau* mais sans l'article *le*. La partie gauche du tableau renseigne un élément crucial : les résultats du modèle pour des EMM non vues dans les données d'apprentissage.

Comme pour nombre de tâches de linguistique computationnelle et de recherche d'information (*Information retrieval*), les résultats mentionnent la précision (P), le rappel (R) et le score F1. La précision est la proportion d'items pertinents parmi l'ensemble des résultats (ici : s'agit-il bien d'une expression verbale lorsque les résultats l'affirment ?) ; le rappel, quant à lui, est la proportion d'items pertinents parmi l'ensemble des items pertinents (ici : parmi toutes les expressions verbales qui devaient être extraites, quel est le pourcentage d'expressions verbales effectivement extraites dans les résultats ?). Le score F1, quant à lui, est la moyenne harmonique entre la précision et le rappel, ce qui permet de vérifier si la méthodologie offre un bon compromis entre la précision (dit-on à chaque fois la vérité ?) et le rappel (dit-on toute la vérité ?).

Dans l'interprétation des données du Tableau 1, il convient de souligner deux écueils de l'extraction automatique en phraséologie : une marge d'erreur ou de subjectivité est toujours présente dans les données de référence (*gold set*); d'autre part, lors des tâches partagées Parseme, les modèles doivent non seulement extraire les expressions verbales, mais les annoter dans le corpus étiqueté. Pour ce dernier point, des erreurs d'étiquetage dans les données de référence (un verbe non reconnu comme tel, des compléments erronément attachés au verbe) se retrouveront également dans les résultats de l'extraction phraséologique.

À titre d'exemple, le Tableau 2 ci-dessous présente une phrase complète traitée par le modèle *HMSid*, dans le format requis par la tâche partagée.

Tableau 2. – Exemple d'extraction phraséologique par le modèle *HMSid*

# source_sent_id = http://hdl.handle.net/11234/1-3105 UD_French-GSD/fr_gsd-ud-train.conllu fr-ud-train_00318										
# text = En 1985, la Constitution des Malouines est entrée en vigueur réduisant grandement le pouvoir du gouverneur, et rendant le gouvernement plus responsable devant le Conseil exécutif et créant un nouveau poste de « chief executive », auquel de nombreux pouvoirs du gouverneur ont été délégués.										
1	En	en	ADP	_	_	2	case	_	_	*
2	1985	1985	NUM	_	_	10	obl:mod	_	SpaceAfter=No	*
3	,	,	PUNCT	_	_	2	punct	_	_	*
4	la	le	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	5	det	_	_	*
5	Constitution	Constitution	PROPN	_	_	10	nsubj	_	_	*
6	de	de	ADP	_	_	8	case	_	_	*
7	les	le	DET	_	Definite=Def|Number=Plur|PronType=Art	8	det	_	_	*
8	Malouines	Malouines	PROPN	_	_	5	nmod	_	_	*
9	est	être	AUX	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	10	aux:tense	_	_	*
10	entrée	entrer	VERB	_	Gender=Masc|Number=Sing|Tense=Past|Typo=Yes|VerbForm=Part	0	root	_	CorrectForm=entrée	1:VID
11	en	en	ADP	_	_	10	iobj	_	EXTPOS=ADV|Type=MWE	1
12	vigueur	vigueur	NOUN	_	Gender=Fem|Number=Sing	11	fixed	_	_	1

Dans le Tableau 2, une phrase de la tâche partagée Parseme 2020 pour le français a été reprise en partie. Les modèles étaient censés y extraire l'expression *entrer en vigueur*, qui devait être étiquetée avec le label VID (*verbal idiom*, expression verbale idiomatique).

L'algorithme extrait toutes les formes verbales : dans le cas d'*entrer en vigueur*, la ligne 10 du tableau renseigne à la fois la forme *entrée* à la colonne 1, le lemme *entrer* (colonne 2) et la partie du discours (verbe, colonne 3). À la ligne 11, *en* est reliée au verbe *entrer*, car l'identifiant 10, qui correspond au verbe *entrer*, est renseigné comme tête grammaticale à la ligne 11, colonne 6 : *en* dépend de *entrer* ; de même, dans la logique de l'étiqueteur automatique fourni par Parseme (UDPipe)[4], *vigueur* dépend de *en*. L'exemple met donc en exergue la complexité de l'extraction phraséologique, car elle est liée à l'étiqueteur, ici UDPipe, qui

4. https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html

eût été mieux inspiré, pour cette phrase, de faire dépendre directement *en vigueur* de *entrer*. Quoi qu'il en soit, le modèle *HMSid* extrait chaque forme verbale du corpus à tester, ainsi que tous ses compléments et les compléments de ces derniers. De cette manière, dans le cas de cette phrase, *entrer en vigueur* se retrouve sur le banc de test, et est soumis au *score-cpr*. Ce dernier est mesuré pour la forme (*entrée en vigueur*) et pour le lemme (*entrer en vigueur*). Si l'un des deux scores est supérieur à 0,45, le modèle assigne le label VID (expression verbale idiomatique). En l'occurrence, le *score-cpr* atteint 0,99 pour *entrée en vigueur* et 0,96 pour *entrer en vigueur*.

Comme le montre le Tableau 1, le modèle *HMSid* a obtenu un score honorable (au vu des scores généralement atteints en linguistique computationnelle) pour l'extraction globale des expressions verbales lors de l'expérience : la précision atteint 66,4 %, le rappel 67,81 % et le score F1 67,10 %. En extraction phraséologique automatisée, le principal défi consiste à reconnaître des constructions totalement absentes des données d'apprentissage. Contrairement aux éditions précédentes, la tâche partagée Parseme 2020 a dès lors décidé d'ajouter ce point aux résultats obtenus par les divers modèles. Étant donné qu'il s'agit d'unités phraséologiques totalement absentes du corpus d'apprentissage, il est d'autant plus difficile d'aboutir à des résultats élevés. Les scores du Tableau 1 (36,49 %) indiquent toutefois un score prometteur, qui est encore susceptible d'amélioration.

Le modèle *HMSid* a obtenu la 5e place globale pour le français dans la section ouverte et la 4e place pour les EMM non vues lors de Parseme 2020. Les meilleurs scores enregistrés pour le français dans la section ouverte de Parseme 2020 étaient nettement supérieurs (un score F1 de 48,01 pour les expressions non vues et 86,39 comme score global pour les tokens). Notons toutefois que la section ouverte permettait l'accès à toute ressource externe (par exemple les dictionnaires ou listes d'expressions), alors que HSMid était uniquement basé sur l'extraction algorithmique à partir d'un seul corpus.

En concertation avec les organisateurs, nous avons ensuite proposé une version adaptée du modèle, *HMSid2*, basée uniquement sur les données d'apprentissage et dès lors classée dans la section dite fermée. Contrairement à la version précédente (qui utilisait le corpus externe WaCky), nous nous sommes cette fois limités aux données d'apprentissage et au corpus extrait de Wikipédia qui avait servi de base à la sélection des données. Les résultats obtenus à l'aide du nouveau modèle *HMSid2* sont présentés dans le Tableau 3.

Tableau 3. – Scores obtenus par le modèle *HMSid2* (Parseme 2020)

EMM non vues				Score global EMM				Score global (tokens)			
P	R	F1	Rang	P	R	F1	Rang	P	R	F1	Rang
32.53	49,33	**39,21**	1	68,90	72,04	**70,43**	2	71,10	72,63	**71,86**	2

Comme l'illustre le Tableau 3, le nouveau modèle HMSid2, proposé en concertation avec les organisateurs de Parseme 2020 après la clôture des résultats officiels, est nettement supérieur au précédent : les résultats globaux atteignent un score F1 de 71,86 % (avec une précision et un rappel tous deux supérieurs à 70 %), ce qui correspond (sur 3 systèmes proposés pour le français) au rang 2 de la section fermée (sans ressources externes) et au rang 1 pour les EMM non vues.

Il convient de souligner que les scores F1 obtenus par HMSid et HMSid2 (Tableaux 1 et 3) pour les expressions non-vues (soit 36,49 et 39,21) sont assez élevés pour un système traditionnel basé sur la linguistique de corpus, sans apprentissage profond. La moyenne du meilleur système pour toutes les langues (expressions non-vues) est d'ailleurs 38,53 (modèle MTLB-STRUCT). Ce modèle obtient la première place, toutes langues confondues, dans la section ouverte. Son score pour le français (expressions non-vues) est un F1 de 42,33. La différence avec HMSid2 (39,21) n'est donc pas énorme, alors que MTLB-STRUCT est un modèle complexe d'apprentissage profond, basé sur des ressources externes.

Notons que, si le modèle HMSid2 (basé sur le corpus Wikipédia) obtient des résultats supérieurs à ceux de HMSid (avec le corpus WaCky), c'est tout simplement parce que les expressions verbales ont été choisies par les organisateurs dans le corpus Wikipédia. Si, d'aventure, telle expression verbale ne comporte aucune ou très peu d'occurrences dans le corpus WaCky, le score statistique est égal à 0. Par contre, toute expression choisie par les organisateurs figurait nécessairement dans le corpus Wikipédia et pouvait donc être évaluée par le score statistique. Ceci met également en lumière un défaut bien connu des corpus par rapport à l'intelligence artificielle : les corpus ne peuvent révéler que ce qu'ils contiennent, alors que l'IA permet d'extrapoler à partir de données d'apprentissage, même si l'expression en question est absente du corpus.

Les résultats de cette expérience montrent qu'un modèle statistique simple (basé sur les clusters de mots, mesurés par le score-*cpr*) permet, dans une large mesure, d'extraire les unités phraséologiques en se fondant sur un vaste corpus. Certes, les modèles peuvent toujours être améliorés, surtout pour les EMM non vues, mais l'interprétation des résultats doit également tenir compte des erreurs d'étiquetage présentes dans les données d'apprentissage, et d'une certaine part de subjectivité inhérente à toute décision prise par les locuteurs natifs quant au caractère phraséologique des constructions.

À l'instar de l'ensemble des résultats proposés lors de Parseme 2020, les données résumées plus haut confirment que la phraséologie, définie comme l'ensemble des constructions figées et/ou idiomatiques, repose largement sur des associations statistiques.

Dans cette contribution, nous nous proposons d'envisager l'algorithmie au-delà de la simple extraction phraséologique. Notre hypothèse de travail est en effet que la phraséologie correspond à une association sémantique d'un type particulier, mais qui s'inscrit dans les associations sémantiques générales : sur

le plan cognitif, pour utiliser la terminologie de Langacker (2008), les constructions morphologiques (les mots), syntaxiques ou autres (dont la phraséologie) correspondent toutes à des degrés divers d'ancrage cognitif (*entrenchment*), dans un vaste réseau. Notre hypothèse est que l'algorithmie permet d'étudier divers aspects de ce réseau, dont la phraséologie.

3. Expérience d'extraction sémantique par apprentissage profond

Le modèle linguistique *BERT* (Devlin *et al.* 2019), ainsi que les modèles dérivés de ce dernier[5] représentent une avancée majeure dans les recherches récentes de linguistique computationnelle.

Il s'agit de modèles linguistiques pré-entraînés selon l'architecture Transformer, qui tient compte de manière détaillée du contexte à gauche et à droite de chaque mot, et donne lieu à un apprentissage par masques linguistiques (en supprimant 15 % des mots de chaque phrase). De cette manière, à partir d'un vaste corpus linguistique, chaque mot (*token*) est représenté par une suite de nombres réels (environ 500 dans la plupart des modèles).

Ces nombres constituent pour chaque mot un vecteur, qui représente ici simplement la liste de tous ces nombres ; ces derniers sont obtenus à l'issue de procédures complexes de plongement (*embedding*), qui permettent de réduire la taille des vecteurs à une taille identique pour chaque mot. L'on peut donc considérer que le modèle pré-entraîné constitue la signature unique de chaque mot dans le corpus, en tenant compte de la plupart des contextes récurrents à gauche et à droite.

Les contextes récurrents sont calculés selon la probabilité de trouver tel mot à gauche ou à droite, dans une certaine fenêtre : par exemple *couper* dans le contexte de *bois*. En comparant le vecteur d'un mot A avec celui d'un mot B, notamment via la similarité cosinus entre les deux vecteurs, on peut estimer la proximité sémantique des mots A et B.

L'architecture Transformer permet d'aller plus loin qu'une simple comparaison entre les vecteurs initiaux, par l'opération d'affinage ou ajustement (*fine-tuning*). À partir d'un nouveau jeu de données, les modèles Transformer pré-entraînés peuvent être téléchargés et soumis à un nouvel ajustement, par apprentissage profond. En quelque sorte, le modèle Transformer pré-entraîné peut être comparé à un dictionnaire de collocations, utilisé pour un nouveau jeu de données linguistiques.

5. Ces divers modèles peuvent être téléchargés à l'adresse https://huggingface.co/transformers/

Le modèle *BERT* et ses dérivés ont permis d'améliorer les résultats obtenus pour nombre de tâches automatisées de la linguistique computationnelle, mais son utilisation en matière de représentation du sens continue de faire débat.

Ainsi, Mickus *et al.* (2020) ont vérifié par plusieurs expériences si le modèle *BERT* offre bel et bien une représentation cohérente des liens de signification en livrant, pour des mots similaires, des résultats semblables. Si la représentation sémantique offerte par *BERT* est cohérente, elle doit en effet assigner les mêmes régions de l'espace sémantique (sous la forme de vecteurs proches) à des paires de mots similaires. Or, l'étude de Mickus *et al.* (2020) indique, au contraire, que la conception même du modèle *BERT* obscurcit la sémantique distributionnelle, par des résultats souvent erronés. Selon les auteurs, ceci tient essentiellement à la fonction *NSP (Next Sentence Prediction)* intégrée dans la structure du modèle : ce dernier vise notamment à prédire quelle est la phrase la plus probable qui va suivre une phrase donnée. Or, ceci affecte la représentation sémantique, en introduisant des résultats différents entre les phrases premières et les phrases secondes. Au bout du compte, la représentation sémantique de la langue offerte par le modèle *BERT* serait en partie incohérente.

Au contraire, l'étude de Karidi *et al.* (2021) semble indiquer l'inverse : en utilisant une technique novatrice, le *MaPP (Masked Pseudoword Probing)*, les auteurs se sont intéressés à certains mots anglais polysémiques, dont le sens est nettement marqué par le contexte (par exemple la préposition anglaise *on* au sens locatif, comme dans *on the table,* par opposition au sens temporel : *on Monday*). Leur technique a permis d'établir une très grande régularité sémantique dans le modèle *BERT* : les sens différents correspondent bel et bien à des régions distinctes de l'espace sémantique représenté par les vecteurs. En résumé, l'utilisation des modèles de type *BERT* à des fins sémantiques reste controversée.

Lors du colloque international Coling 2020 s'est également tenue la cinquième édition de la tâche partagée CogALex (Xiang *et al.* 2020). Elle consistait en l'extraction spécifique de synonymes, antonymes et hyperonymes à partir de données d'apprentissage en allemand, anglais, et chinois (mandarin). L'italien a été ajouté lors de la phase de test, mais ne comportait pas de données d'apprentissage.

Les quatre étiquettes possibles étaient : SYN (synonymes), ANT (antonymes), HYP (hyperonymes) et RANDOM (aléatoire, c'est-à-dire absence d'association sémantique). Le corpus d'apprentissage était étiqueté de la sorte, pour des paires de mots ou des paires de mots composés. Le Tableau 3 reprend, à titre d'exemple, le début des données d'apprentissage pour l'anglais. Le français n'était malheureusement pas repris parmi les langues de CogALex 20.

Tableau 4. – Extrait des données d'apprentissage pour l'anglais (CogALex 2020)

construe	Mingle	RANDOM
leaflet	Book	ANT
ink	Quiet	RANDOM
citation	Source	SYN
disjoint	Mend	ANT
neural	Permissive	RANDOM
disciplinary	unacceptable	RANDOM
musician	Audiophobe	ANT
wink	Impress	RANDOM
geologic	Metallic	SYN
disclose	Show	SYN
inconsolable	Annual	RANDOM
left	Right	ANT
tour	Journey	HYP
representative	well_prepared	RANDOM

Le meilleur modèle proposé lors de la tâche partagée CogALex 2020, *Text2TCS* (Wachowiak *et al.* 2020) était basé sur un modèle proche de *BERT* : *XLM-RoBERTa*. Ses principaux résultats sont présentés dans le Tableau 5.

Tableau 5. – Résultats (scores F1) du modèle *Text2TCS* lors de CogALex 2020

Langue	SYN	HYP	ANT	Moyenne
anglais	0,473	0,483	0,587	0,517
chinois	0,849	0,876	0,914	0,881
allemand	0,427	0,535	0,534	0,500

Le Tableau 5 le montre : les résultats d'un modèle d'apprentissage profond basé sur une architecture Transformer sont prometteurs : le score F1 se situe aux alentours de 50 %, ce qui représente déjà un exploit en extraction automatique des synonymes, antonymes et hyperonymes. Le chinois constitue en outre une exception notoire, avec des résultats impressionnants qui posent question : s'agit-il d'un artefact de la procédure ou certains facteurs liés à la langue chinoise interviennent-ils ?

Nous avons mené une nouvelle expérience avec les données de CogALex 2020, afin de vérifier ces résultats et de tenter d'éclairer quelque peu le débat théorique autour des modèles de type *BERT*.

Dans notre expérience, la méthodologie utilisée suit la procédure habituelle lors de l'ajustement des modèles pré-entraînés. Nous avons toutefois sélectionné ces modèles, tous dérivés de *BERT*, en fonction de leur adaptation contextuelle optimale selon la langue considérée.

Pour l'anglais, nous avons ainsi retenu le modèle pré-entraîné *dbmdz/electra-large-discriminator-finetuned-conll03-english*[6]. À titre indicatif, ses principales caractéristiques sont reprises dans le Tableau 6.

Tableau 6. – Caractéristiques du modèle pré-entraîné dbmdz/electra-large-discriminator-finetuned-conll03-english

anglais: modèle pré-entraîné	
Architecture	Transformer
nombre de couches	16
nombre de couches cachées	24
taille vectorielle maximale	512
fonction d'activation	gelu
taille du vocabulaire	30522

Pour les autres langues, nous avons utilisé des modèles dont la structure est fort proche : pour l'allemand, le modèle pré-entraîné *german-nlp-group/electra-base-german-uncased*[7] et pour le chinois le modèle *BERT-base-chinese*[8].

Notons au passage que nous avons utilisé un modèle *BERT* pour le chinois, à défaut d'un meilleur modèle disponible ; pour l'allemand et l'anglais, il s'agit d'un modèle *Electra* (*Efficiently Learning an Encoder that Classifies Token Replacement Accurately*). Ce dernier représente une légère amélioration par rapport au modèle *BERT* initial, grâce à un système de masques linguistiques plus efficace (certains mots, arbitrairement remplacés par un vide dans *BERT* afin d'améliorer l'apprentissage, sont ici remplacés par des mots plausibles, générés de manière statistique). Le meilleur système présenté lors de CogALex 2020, *Text2TCS* (Tableau 5) était basé sur un autre modèle proche de *BERT* : *XLM-RoBERTa*.

Dans les grandes lignes, le processus d'apprentissage profond à partir d'un modèle pré-entraîné se déroule comme suit, dans le cas d'une classification

6. https://huggingface.co/dbmdz
7. https://huggingface.co/german-nlp-group/electra-base-german-uncased/tree/main
8. https://huggingface.co/BERT-base-chinese

catégorielle. Dans le cas présent, les 4 catégories sont SYN (synonyme), ANT (antonyme), HYP (hyperonyme) et RANDOM (aléatoire). Pour chaque paire de mots des données d'apprentissage, une régression statistique complexe est appliquée afin d'obtenir la meilleure pondération possible entre les données du modèle pré-entraîné. Celui-ci ne comporte pas moins de 40 couches (qui représentent chacune une description numérique des contextes et associations de chaque mot), pour une taille vectorielle maximale de 512. Chacun des 30522 mots du corpus anglais est donc représenté par un vecteur de 512 nombres réels. Ces 512 nombres réels sont présents dans les 40 couches du modèle neuronal entièrement interconnecté, ce qui constitue un gigantesque tableau de 512 x 40 = 20 480 neurones. Ces derniers sont manipulés dans un seul et même tableau multidimensionnel baptisé *tenseur* par des fonctions mathématiques complexes, de type probabiliste. Un optimiseur d'apprentissage (*Adam*) tente de minimiser le nombre de calculs aléatoires, et une fonction d'activation des neurones (ici la fonction *gelu*) tente à chaque *époque* (une époque correspond à une traversée complète du corpus) d'arriver à la meilleure adéquation possible avec les données de référence, à partir d'un échantillon.

La manipulation de *tenseurs* gigantesques par des fonctions mathématiques complexes n'est pas à la portée d'un ordinateur de base. Cependant, contrairement aux idées reçues, la plupart des expériences linguistiques d'apprentissage profond peuvent s'effectuer en ligne grâce aux plateformes gratuites dédiées[9].

Pour cette expérience, nous avons utilisé l'outil d'apprentissage automatique TensorFlow[10], invoqué en langage Python via les librairies Keras et Ktrain. L'ajustement (*fine-tuning*) d'un modèle neuronal pré-entraîné est relativement simple à l'aide de TensorFlow et du langage Python, grâce à l'emploi de fonctions intégrées dans la librairie Transformer[11].

L'apprentissage profond a permis un ajustement optimal à l'issue de 6 époques seulement, soit en une trentaine de minutes à peine. Les résultats obtenus lors de l'apprentissage aboutissaient à un score F1 moyen de 0,79 pour l'anglais, 0,67 pour l'allemand et 0,88 pour le chinois. Le modèle a ensuite été testé à partir du jeu de données et du script d'évaluation en Python fournis par CogALex 2020. Ces résultats sont présentés dans le Tableau 7.

---

9. Pour l'apprentissage profond, il est nécessaire d'optimiser la carte graphique (GPU) de l'ordinateur en plus de son microprocesseur (CPU). Même après optimisation, les modèles Transformer nécessitent une capacité graphique telle qu'une plateforme dédiée est nécessaire, par exemple Google Colab (https://colab.research.google.com) ou Kaggle (https://www.kaggle.com). Pour les modèles gigantesques, il faut recourir à un TPU (*tensor processing unit*), qui doit toutefois être programmé spécifiquement.
10. https://www.tensorflow.org/?hl=fr
11. Nous avons utilisé, dans la librairie Transformer de Python, les fonctions *AutoTokenizer*, et *AutoModelForSequenceClassification*. Les meilleurs résultats ont été obtenus avec les hyperparamètres suivants : *Learning rate* = 2e-5, *Epochs* = 6, Batch size = 48.

Tableau 7. – Résultats (scores F1) du modèle *HSemId2* évalués par CogALex2020

Langue	SYN	HYP	ANT	Moyenne
anglais	0,627	0,584	0,779	0,668
chinois	0,852	0,869	0,919	0,882
allemand	0,486	0,579	0,728	0,600

Le Tableau 7 reprend les résultats obtenus par notre modèle, baptisé *HSemId2*, basé sur une architecture Transformer. Ils confirment et sont même supérieurs à ceux du meilleur modèle lors de CogALex 2020 (Tableau 5) ; le score F1 est légèrement supérieur pour le chinois, et nettement supérieur pour l'anglais et l'allemand (respectivement 0,668 et 0,600 pour *HSemId2*, contre 0,517 et 0,500 pour le modèle *Text2TCS*).

Les résultats obtenus sont particulièrement élevés pour le chinois, comme c'était le cas avec le modèle *Text2TCS*. La répétition de tels résultats par le biais de deux expériences distinctes suggère que ces scores élevés sont liés à des particularités de la langue chinoise et non à des hasards de manipulation des données. Notre hypothèse pour expliquer les scores qui restent nettement plus élevés en chinois est la suivante : l'architecture Transformer repose en large partie sur un « segmenteur » (*tokenizer*) propre à chaque modèle. Dans les modèles Transformer, la taille du vocabulaire doit rester dans les limites du raisonnable (par exemple, au Tableau 8 : 30522 mots pour l'anglais), afin de ne pas augmenter la taille des listes multidimensionnelles ou *tenseurs*. Lorsque des mots absents du vocabulaire sont soumis au modèle, celui-ci les décompose et tente de les traiter en unités de sens. Il va sans dire qu'une telle stratégie est souvent problématique pour les langues européennes, étant donné les aléas des constructions morphologiques. Les mots chinois (dont la plupart comportent deux ou trois caractères, parfois un seul), en revanche, se prêtent mieux à une segmentation de ce type, car la langue est bien plus isolante que les langues européennes. *Université*, par exemple, se dit en mandarin 大学 (*dàxué*), qui peut à son tour se décomposer en 大 (*dà*, grand, supérieur) et 学 (*xué,* apprendre, apprentissage) : l'université est l'enseignement supérieur et la segmentation du mot garde bien une cohérence sémantique.

D'un point de vue théorique, les réserves mentionnées plus haut à propos de la validité de la représentation sémantique des modèles de type *BERT* (Mickus *et al.* 2020) ne sont que partiellement confirmées par les résultats de notre expérience (Tableau 7). La distinction subtile des synonymes et des antonymes, par exemple, est une tâche complexe pour la linguistique computationnelle. Prenons l'exemple de *gauche*, qui peut être considéré comme un antonyme de *droite* : dans

la plupart des langues, leurs contextes seront très proches (on dit par exemple *tourner à gauche, tourner à droite*). La moyenne des scores F1 obtenus via un modèle de type *BERT* pour l'anglais et l'allemand (Tableau 7) est relativement faible (respectivement 0,668 et 0,600), mais le score est nettement plus élevé pour le chinois (0,882) et ces chiffres corroborent dans les grandes lignes les résultats du meilleur modèle (*Text2TCS*) lors de CogALex 2020 (cf. Tableau 5).

Une autre hypothèse permettant peut-être d'éclairer ce débat théorique concerne à nouveau la phraséologie, au sens général de toutes les associations idiomatiques ou partiellement idiomatiques. Comme nous l'avons signalé dans la section 2, le meilleur système d'extraction phraséologique lors de Parseme 2020 était également basé sur l'apprentissage profond et sur le modèle *BERT*. Cette méthodologie semble bien fonctionner pour l'extraction de la phraséologie. Or, comme nous l'avons vu, le sens distributionnel et le sens phraséologique présentent de nombreux points d'interaction. Il est donc possible que les résultats pertinents obtenus pour l'extraction sémantique (Tableau 7) soient en partie liés à une extraction phraséologique. Ceci pourrait expliquer aussi les meilleurs résultats obtenus pour le chinois. En effet, les collocations et autres associations idiomatiques sont souvent imbriquées, en chinois, dans le lexique. Par exemple, enseignant, professeur se dit en mandarin 老师 (*lǎoshī*). Il s'agit d'un mot du lexique, au sens où nous l'entendons dans les langues européennes, mais ce mot se décompose en 老 (*lǎo*), vieux, et 师 (*shī*), maître et pourrait dès lors être vu comme une collocation liée à la culture chinoise : *un vieux maître*.

Notre hypothèse de travail est donc que les bons résultats obtenus par les modèles de type *BERT* pour l'extraction sémantique sont (en partie) liés à l'extraction phraséologique. Cette hypothèse est difficile à vérifier, car les modèles d'apprentissage profond fonctionnent en grande partie comme une boîte noire : les algorithmes lancent des approximations statistiques de manière aléatoire, qui sont contrôlées par des fonctions mathématiques. Au terme de l'apprentissage, le modèle fournit toute une série de pondérations mathématiques dont on peut mesurer l'efficacité mais difficilement expliquer la logique précise.

Si notre hypothèse d'une influence phraséologique sur l'extraction sémantique dans les modèles *BERT* est exacte, nous pouvons prédire qu'un modèle tel que celui que nous avons proposé au Tableau 7, pour les synonymes, antonymes et hyperonymes, devrait fournir des résultats encore meilleurs pour l'extraction phraséologique. Pour tenter d'éclaircir ce point, nous proposons un modèle identique à celui du Tableau 7 et l'appliquons à l'extraction phraséologique.

4. Expérience d'extraction phraséologique par apprentissage profond

À des fins d'extraction phraséologique, nous utilisons dans cette expérience, limitée à l'anglais, un modèle (*HSemId3)* en tout point semblable à *HSemId2* (qui visait l'extraction sémantique).

Comme pour *HSemId2*, il s'agit d'une technique d'ajustement du même modèle pré-entraîné, dérivé de *BERT*[12]. Dans le cas du modèle *HSemId2*, les quatre catégories à identifier étaient : SYN (synonymes), ANT (antonymes), HYP (hyperonymes) et RANDOM (associations aléatoires). Le Tableau 7 (plus haut), généré par le script d'évaluation automatique de CogALex 2020, ne reprenait que les résultats pour les synonymes, antonymes et hyperonymes (sans la catégorie RANDOM). Afin de faciliter la comparaison avec *HSemId3*, le Tableau 8 mentionne les meilleurs résultats obtenus pour *HSemId2* lors de l'entraînement du modèle, pour les 4 catégories.

Tableau 8. – Résultats des données d'entraînement (anglais) du modèle *HSemId2* pour l'extraction sémantique (données de CogALex20)

Catégories	Précision	Rappel	Score F1
SYN	0,49	0,42	0,45
HYP	0,46	0,55	0,50
ANT	0,71	0,69	0,70
RANDOM	0,86	0,85	0,85

Les résultats présentés au Tableau 8 pour l'extraction sémantique en anglais correspondent donc aux meilleurs scores obtenus lors de l'entraînement du modèle avec les données de CogALex20 ; notons au passage que les résultats définitifs calculés à partir des données tests par le programme automatisé de CogALex20 étaient supérieurs (Tableau 7). Un point important concerne aussi les scores obtenus pour la catégorie aléatoire (RANDOM) : en extraction sémantique, la différence entre synonymes, antonymes et hyperonymes est parfois très subtile, tandis que l'absence de toute relation sémantique, dénotée par la catégorie RANDOM, est plus facile à reconnaître par le modèle, avec un score F1 de 0,85 dans notre expérience.

Notre nouveau modèle adapté à l'extraction phraséologique, *HSemdId3*, comporte également 4 catégories, afin de faciliter la comparaison avec *HSemId2*. Nous avons utilisé, pour la langue anglaise, trois catégories générales de la phraséologie : IDIOM (expression idiomatique, par exemple *spill the beans*, cracher

12. Il s'agit du modèle mentionné dans la section précédente pour l'anglais : dbmdz/electra-large-discriminator-finetuned-conll03-english

le morceau, révéler un secret), COLLOCATION (collocation, par exemple *harsh criticism*, de sévères critiques), FORMULA (formule communicative, comme dans *how are you doing, comment allez-vous*). Comme équivalent de la catégorie RANDOM (association aléatoire) pour l'extraction sémantique, nous avons ajouté une quatrième catégorie pour l'extraction phraséologique : FREE (combinaison libre, par exemple *and then not the*, et alors pas le).

La compilation d'une base de données phraséologiques est complexe, car il y a rarement unanimité parmi les locuteurs natifs, surtout dans le cas de constructions semi-idiomatiques. Nous avons dès lors choisi de constituer nous-même une base de données phraséologiques partielle de 1500 entrées pour chacune des 4 catégories IDIOM, COLLOCATION, FORMULA, FREE. Pour les trois premières catégories, nous nous sommes basé sur les listes fournies par les dictionnaires ; pour la catégorie *FREE*, nous avons extrait d'un corpus (ukWaC, Baroni *et al.* 2009) les combinaisons de 2, 3, 4 et 5 mots de fréquence élevée dont la nature n'était clairement pas phraséologique mais purement grammaticale ou communicative (par exemple *and then he, and although it is*). Parmi les 9000 entrées de la base de données, deux tiers ont été utilisées comme données d'entraînement et un tiers comme données test. Les meilleurs résultats obtenus lors de l'entraînement[13] figurent dans le Tableau 9.

Tableau 9. – Résultats des données d'entraînement (anglais) du modèle *HSemId3* pour l'extraction phraséologique

Catégories	Précision	Rappel	Score F1
IDIOM	0,86	0,86	0,86
COLLOCATION	0,86	0,94	0,88
FORMULA	0,87	0,94	0,90
FREE	0,98	0,96	0,97

La comparaison est aisée entre les Tableaux 8 et 9 : pour une même architecture basée sur l'ajustement (*fine-tuning*) du même modèle pré-entraîné, et des paramètres identiques, les résultats sont nettement meilleurs en extraction phraséologique.

Rappelons que les résultats mesurés au Tableau 8 avec les données de CogALex20 sont toutefois supérieurs à ceux obtenus par le meilleur système lors de la tâche partagée ; ces résultats restent faibles pour l'anglais et l'allemand mais sont élevés pour le chinois.

13. Comme indiqué plus haut, les données techniques de l'entraînement du modèle *HSemId3* (extraction phraséologique) étaient rigoureusement identiques à celles de *HSemId2* (extraction sémantique), cf. note 11.

Les résultats très élevés mentionnés au Tableau 9 pour l'extraction phraséologique confirment que les modèles de type *BERT* sont très efficaces pour l'extraction des associations idiomatiques, ainsi que la tâche partagée Parseme 2020 (section 2) l'avait déjà démontré.

5. Conclusions

Comme nous l'avons relevé dans l'introduction, le sens phraséologique et le sens distributionnel paraissent a priori contradictoires : *lever le pied* aura, dans certains contextes, un sens littéral (compositionnel) et distributionnel (qui répond aux associations sémantiques naturelles) : exercer un mouvement du pied vers le haut ; dans d'autres cas, il s'agira d'une unité phraséologique dont le sens est global et idiomatique : ralentir, travailler de manière moins intense.

Les expériences dont nous avons rapporté les résultats contredisent quelque peu cette vision tranchée des domaines distributionnel et phraséologique.

Au point 2, les divers résultats que nous avons obtenus par l'algorithmie lors de la tâche partagée Parseme 2020 confirment avant tout le caractère hautement probabiliste des associations phraséologiques. Les algorithmes utilisés se basent uniquement sur les récurrences statistiques et permettent d'obtenir une précision et un rappel élevés, surtout si l'on tient compte de la subjectivité relative que comporteront toujours les données de référence. D'année en année, les résultats obtenus par les programmes d'extraction phraséologique s'améliorent et nos propres résultats obtenus pour le français, avec un score F1 supérieur à 70 %, le confirment.

Dans la conception et l'analyse des données du modèle phraséologique, nous avons rencontré à maintes reprises la problématique du sens : démêler l'écheveau du sens, entre phraséologie et sens distributionnel s'avère souvent malaisé.

L'introduction de l'architecture Transformer en 2019 a révolutionné l'apprentissage profond du langage par les modèles pré-entraînés. Cette méthodologie, par sa nature même, correspond à une vision plus phraséologique de la langue, car elle repose sur un apprentissage exhaustif des contextes à gauche et à droite de chaque mot, ainsi que sur la probabilité de retrouver les mots masqués. Nous avons ainsi proposé une nouvelle expérience à partir du jeu de données de CogALex 2020, et l'avons évaluée à l'aide du programme fourni par la tâche partagée. Cette méthodologie a fourni des résultats nettement supérieurs aux meilleurs scores obtenus lors de la tâche partagée CogALex 2020. Nous avons ensuite testé, pour l'anglais, l'application de ce modèle sémantique à l'extraction phraséologique. Il s'est avéré que les résultats de notre modèle étaient encore nettement supérieurs pour l'extraction phrasésologique. Ceci confirme que l'apprentissage profond basé sur l'architecture Transformer permet effectivement de démêler quelque peu l'écheveau sémantique, mais sans doute en partie par le détour de la phraséologie.

Bibliographie

BALLY Ch. (1921). *Traité de stylistique française*. Heidelberg : Carl Winter's Universitätsbuchandlung.

BARON M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation* 43, 209-226.

BARONI M. (2013). Composition in distributional semantics. *Language and Linguistics Compass* 7, 511-522.

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

COLSON J.-P. (2017). The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. In : R. Mitkov (ed), *Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596*. Cham, Springer International Publishing, 16-28.

COLSON J.-P. (2018a). Les traces du figement dans les corpus linguistiques : une étude de cas. *Le français moderne* 86, 129-145.

COLSON J.-P. (2018b). From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin. In : SAVARY, A., RAMISCH, C., HWANG, J. D., SCHNEIDER, N., ANDERSEN, M., PRADHAN, S., PETRUCK, M. R. L. (eds), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe (New Mexico, USA), Association for Computational Linguistics, 41-50.

COLSON J.-P. (2020a). HMSid and HMSid2 at PARSEME Shared Task 2020: Computational Corpus Linguistics and unseen-in-training MWEs. In : *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. Barcelona, Online, Association for Computational Linguistics, 119-123.

COLSON J.-P. (2020b). Extracting meaning by idiomaticity: Description of the HSemID system at CogALex VI (2020). In : *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*. Barcelona, Online, Association for Computational Linguistics, 54-58.

CORPAS PASTOR G. & COLSON J.-P. (eds) (2020). *Computational Phraseology*. Amsterdam, Philadelphia : John Benjamins.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Association for Computational Linguistics, 4171-4186.

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

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

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

FIRTH J.R. (1957). *A synopsis of linguistic theory, 1930-1955*. Oxford : Blackwell.

HARRIS S. (1954). Distributional structure. *Word* 10, 146-162.

HEYLEN K. & BERTELS A. (2016). Sémantique distributionnelle en linguistique de corpus. *Langages* 201, 51-64.

KARIDI T., ZHO Y., SCHNEIDER, N., ABEND O. & SRIKUMAR V. (2021). Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords. In : *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP21)*, November 7-11. Association for Computational Linguistics, 10300-10313.

LANGACKER R. (2008). *Cognitive Grammar. A Basic Introduction*. Oxford : Oxford University Press.

LAPESA G. & EVERT S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics* 2, 531-545.

LENCI A. (2018). Distributional models of word meaning. *Annual review of Linguistics* 4, 151-171.

MEJRI S., MENESES-LERÍN L. & BUFFARD-MORET B. (éds) (2020). *La phraséologie française en questions.* Paris : Hermann Éditeurs.

MICKUS T., PAPERN D., CONSTANT, M. & VAN DEEMTER K. (2020). What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. In : *Proceedings of the Society for Computation in Linguistics (SCiL)* 2020. New Orleans, Louisiana, January 2-5, 350-361.

MITCHELL J. & LAPATA M. (2010). Composition in distributional models of semantics. *Cognitive science* 34, 1388-429.

NELSON D.L., McEVOY C. L. & SCHREIBER T.A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/FreeAssociation/.

WACHOWIAK L., LANG, C., HEINISCH B. & GROMANN D. (2020). CogALex-VI Shared Task: Transrelation – A Robust Multilingual Language Model for Multilingual Relation Identification. In : *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. Barcelona, Online : Association for Computational Linguistics, 59-64.

XIANG R., CHERSONI E., IACOPONI L. & SANTUS E. (2020). The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations. *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. Online, Association for Computational Linguistics, 46-53.