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

De l'interprétabilité des dimensions à l'interprétabilité du vecteur : parcimonie et stabilité

Simon Guillot[1] Thibault Prouteau[1] Nicolas Dugué[1]
(1) LIUM, Le Mans, France
prenom.nom@univ-lemans.fr

RÉSUMÉ

Les modèles d'apprentissage de plongements parcimonieux (SPINE, SINr) ont pour objectif de produire un espace dont les dimensions peuvent être interprétées. Ces modèles visent des cas d'application critiques du traitement de la langue naturelle (*e.g.* usages médicaux ou judiciaires) et une utilisation des représentations dans le cadre des humanités numériques. Nous proposons de considérer non plus seulement l'interprétabilité des dimensions de l'espace de description, mais celle des vecteurs de mots en eux-mêmes. Pour cela, nous introduisons un cadre d'évaluation incluant le critère de stabilité, et redéfinissant celui de la parcimonie en accord avec les théories psycholinguistiques. Tout d'abord, les évaluations en stabilité indiquent une faible variabilité sur les modèles considérés. Ensuite, pour redéfinir le critère de parcimonie, nous proposons une méthode d'éparsification des vecteurs de plongements en gardant les composantes les plus fortement activées de chaque vecteur. Il apparaît que pour les deux modèles SPINE et SINr, de bonnes performances en similarité sont permises par des vecteurs avec un très faible nombre de dimensions activées. Ces résultats permettent d'envisager l'interprétabilité de représentations éparses sans remettre en cause les performances.

ABSTRACT

From dimension to vector interpretability : sparseness and stability

Sparse word embeddings models (SPINE, SINr) are designed to embed words in interpretable dimensions. These models are useful for critical downstream tasks in natural language processing (*e.g.* medical or legal NLP), and digital humanities applications. We propose to shift attention from the interpretability of dimensions to the interpretability of word vectors. We thus introduce stability to the interpretability framework, and redefine sparseness. First, stability evaluations show some variability for both models. Then, our sparsifiaction approach redefines the sparseness criterion by keeping only a limited number of components among the strongest in each vector. Both SPINE and SINr show interesting performances on the similarity task with very few activated dimensions. These results are encouraging and pave the way towards intrinsically interpretable word embeddings.

MOTS-CLÉS : Sémantique distributionnelle, traits sémantiques, interprétabilité, plongements.

KEYWORDS: Distributional semantics, semantic features, interpretability, word embeddings.

1 Introduction

L'interprétabilité (Rudin, 2019) telle qu'elle est définie dans la littérature pour les modèles de plongements lexicaux correspond à assurer la possibilité de trouver une cohérence sémantique (ou syntaxique) aux dimensions de l'espace de représentation (Murphy *et al.*, 2012; Senel *et al.*, 2018;

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 83–91.

Subramanian *et al.*, 2018; Prouteau *et al.*, 2022). Cette interprétabilité facilite également la mise en perspective des représentations du lexique avec des conceptions linguistiques de celui-ci. En effet, des dimensions de description sémantiquement cohérentes sont analysables comme traits sémantiques ou sèmes, unités utilisées dans une variété de modèles théoriques de représentation du sens allant de modèles d'inspiration cognitive (Jackendoff, 1983) à des modèles structuralistes ou néosaussuriens (Pottier, 1963; Rastier, 2009). Ces unités de sens varient quant à leur nombre, leur caractère universel ou relatif (à une langue, ou à un corpus), leur statut de primitives sémantiques et leur portée (le référent ou le concept) selon les cadres théoriques (Rastier, 2009).

Les modèles de plongements denses de l'état de l'art (Mikolov *et al.*, 2013; Pennington *et al.*, 2014; Devlin *et al.*, 2018) ont permis des évolutions importantes dans le traitement automatique. Ils consistent à plonger le lexique dans des espaces de représentation denses aux dimensions opaques. Il est possible d'obtenir une compréhension a posteriori de ces modèles par le sondage (Rogers *et al.*, 2021) ou l'analyse des matrices de plongements (Shin *et al.*, 2018). Ces méthodes correspondent à l'*explicabilité* en apprentissage automatique. L'article fondateur de Murphy *et al.* (2012) ouvre la réflexion sur les représentations distributionnelles psycholinguistiquement plausibles. Les auteurs postulent un ensemble de contraintes sur l'espace de représentation : la parcimonie, la positivité, et la performance. La parcimonie est justifiée par la difficulté de couvrir l'ensemble du vocabulaire par un faible nombre de traits partagés. Ainsi, de nombreuses dimensions sont nécessaires, mais seules certaines sont activées pour la description d'un mot. La positivité tient au caractère non économique du stockage d'informations négatives, critère déjà utilisé pour des travaux psychologiquement motivés (Palmer, 1977) sur la factorisation de matrices d'éléments perceptifs (Lee & Seung, 1999). Les travaux sur les modèles de plongement épars ou parcimonieux connaissent des développements avec `SPOWV` (Faruqui *et al.*, 2015), `SPINE` (Subramanian *et al.*, 2018) et `SINr` (Prouteau *et al.*, 2021). Les deux premiers transforment un espace de représentation dense en espace épars, tandis que le troisième construit un espace épars depuis la matrice de cooccurrences.

La positivité et la parcimonie constituent des caractéristiques nécessaires à l'interprétabilité des représentations distributionnelles du lexique. Néanmoins, cette interprétabilité est cantonnée au niveau de la dimension d'un espace de représentation. En effet, les tests d'intrusion qui servent à la caractériser dans Murphy *et al.* (2012); Senel *et al.* (2018); Subramanian *et al.* (2018); Prouteau *et al.* (2022) examinent les dimensions une à une pour vérifier leur cohérence sémantique interne, ainsi que l'activation de la dimension pour un faible nombre d'items lexicaux.

Dans ce travail, nous proposons de concevoir également l'**interprétabilité au niveau du vecteur** de mot et non plus seulement au niveau des dimensions. Cette interprétabilité devient possible si un mot est décrit par une faible quantité de dimensions dont il est possible de faire sens. En effet, d'une part les tâches d'association utilisées pour établir des listes de traits sémantiques comme celles de (Garrard *et al.*, 2001; McRae *et al.*, 2005) indiquent l'ordre de grandeur d'une dizaine de traits par item à décrire en sommant les réponses de leurs annotateurs. D'autre part, (Miller, 1956; Peterson & Peterson, 1959) posent une limite pour la manipulation d'items lexicaux en mémoire de travail cohérente avec cet ordre de grandeur. Nous prenons ainsi le parti de définir cette quantité de descripteurs comme un horizon souhaitable pour les représentations interprétables, dans la mesure où de tels vecteurs restreints en dimensions sont plus manipulables par d'éventuels locuteurs pour en faire sens. Par ailleurs, nous proposons de définir la stabilité comme un critère supplémentaire à l'interprétabilité des vecteurs. Que ce soit dans le cadre des humanités numériques (Hellrich & Hahn, 2016a,b), ou dans des utilisations critiques du traitement automatique de la langue pour des applications juridiques ou médicales (Digan *et al.*, 2020), la reproductibilité des résultats est un critère essentiel. Or, l'entraînement non déterministe des modèles de plongements neuronaux cause

une certaine instabilité dans les représentations et les voisinages, même pour des représentations entraînées avec les mêmes hyperparamètres sur les mêmes données (Pierrejean, 2020). Rudin (2019) encourage à prioriser les approches interprétables sur les approches explicables qui souffrent de cette instabilité, même si la littérature voit émerger de récents efforts pour la création de méthodes d'explicabilité *a posteriori* déterministes (Zafar & Khan, 2021).

Ainsi, dans cet article, nous évaluons la capacité des modèles parcimonieux de l'état de l'art à se conformer aux contraintes d'interprétabilité des vecteurs que nous définissons : la stabilité de ces vecteurs, et la capacité à produire des représentations performantes avec un très faible nombre de dimensions activées, eu égard aux travaux en psycholinguistiques. Nous décrirons d'abord en Section 2 les deux modèles considérés dans nos expérimentations : `SPINE` et `SINr`. Puis nous détaillerons en Section 3 notre cadre expérimental : les corpus utilisés (OANC et BNC) pour apprendre les plongements et les jeux de données pour les évaluer sur des tâches de similarité. Nous introduirons également notre approche d'éparsification des vecteurs pour produire progressivement des modèles qui tendent vers une dizaine de dimensions activées par vecteur. Enfin, nous présenterons en Section 4 les résultats en stabilité et en similarité sur des représentations éparsifiées qui permettent de tendre vers la définition réactualisée de l'interprétabilité que nous proposons.

2 Les modèles interprétables

SPINE. `SPINE` est introduit par Subramanian *et al.* (2018) et permet, à partir d'une représentation dense obtenue avec `Word2Vec` (Mikolov *et al.*, 2013) ou `GloVe` (Pennington *et al.*, 2014), de produire un modèle parcimonieux dans un espace de plus grande taille (*e.g.* 1000 dimensions). Pour cela, `SPINE` est un auto-encodeur dont la couche cachée est en plus grande dimension que l'entrée à reconstruire. De plus, l'auto-encodeur est dit *k-sparse*, l'objectif est donc de n'activer que k neurones dans la couche cachée pour réussir à reconstruire l'entrée. Pour apprendre un tel modèle, trois fonctions de coût sont employées. La fonction de coût de reconstruction (*Reconstruction Loss*) pénalise la mauvaise reconstruction de l'entrée à partir de la représentation parcimonieuse fournie par la couche cachée. Pour imposer l'activation d'un faible nombre de neurones dans la couche cachée et donc la parcimonie de la matrice de plongements, les auteurs introduisent la fonction de coût sur la parcimonie moyenne (*Average Sparsity Loss*) et la fonction de coût sur la parcimonie partielle (*Partial Sparsity Loss*). Ces deux fonctions pénalisent un trop grand nombre d'activations tout en forçant les valeurs d'activation vers 0 ou 1. Le modèle `SPINE` présente plusieurs hyperparamètres : le niveau minimal de parcimonie, le nombre d'époques d'apprentissage et la dimension de la représentation.

SINr. Introduite dans Prouteau *et al.* (2021), `SINr` est une approche basée graphes pour la représentation distributionnelle du lexique. En appliquant une fenêtre glissante sur les phrases du corpus, la méthode extrait un graphe de co-occurrences pondéré (les sommets sont des mots et les poids des arêtes le nombre de co-occurrences observées). Prouteau *et al.* (2021) appliquent ensuite un algorithme de détection de communautés (l'algorithme de *Louvain* (Blondel *et al.*, 2008)) pour extraire des communautés de mots densément connectés donc fréquemment co-occurrents. À partir de cette partition du graphe en communautés, `SINr` calcule la distribution des liens de chaque sommet à travers les communautés détectées. Cela permet ainsi d'avoir une distribution de chaque mot sur les groupes de mots découverts de façon non supervisée par l'algorithme. Les plongements de mots produits sont naturellement parcimonieux, un sommet n'est pas connecté à toutes les communautés.

Le modèle SINr n'a qu'un seul hyperparamètre qui agit sur le nombre de communautés détectées.

3 Cadre expérimental

Modèles. Outre les modèles parcimonieux présentés Section 2, nous utilisons Word2Vec comme contrôle. Nous utilisons *Skipgram with Negative sampling* avec les paramètres décrits dans Levy & Goldberg (2014), la dimension des plongements de mots est de 300 avec une fenêtre de contexte de taille 5. Puisque le nombre de dimensions pour SPINE peut être fixé facilement contrairement à la méthode SINr (il dépend du nombre de communautés détectées), on fixe le nombre dimensions de SPINE en fonction de celles obtenues avec SINr. Les performances de la méthode SINr semblent optimales en réglant l'hyperparamètre agissant sur la détection de communautés à 50, aboutissant à 8454 dimensions pour BNC et 4460 pour OANC, ces nombres sont donc également choisis pour SPINE. Les plongements de mots SPINE sont appris à partir du modèle Word2Vec présenté ci-avant. Le niveau de parcimonie obtenu avec SPINE est peu sensible à l'hyperparamètre censé permettre de le régler. Ainsi, un grand nombre de modèles a été lancé et les résultats sont présentés sur le modèle qui obtient les meilleures performances sur la tâche d'évaluation en similarité avec une parcimonie (1000 époques aboutissant à 95% de parcimonie) permettant au modèle de subir le protocole que nous décrivons ci-après.

Protocole expérimental. Dans ces travaux, nous introduisons un protocole expérimental permettant d'évaluer l'interprétabilité au niveau des vecteurs. Pour cela, nous commençons par considérer le compromis performance-parcimonie. Nous supposons que des vecteurs plus épars sont à la fois plus interprétables et plus plausibles psycholinguistiquement. Pour travailler avec des modèles à la parcimonie contrôlée, nous introduisons notre approche d'éparsification des vecteurs : des représentations de chaque modèle de plongements sont construites en conservant pour chaque vecteur les k composantes ayant les valeurs les plus élevées, en variant k de 250 à 10. Les autres composantes sont fixées à 0. Si ce processus est naturel pour les modèles *a priori* parcimonieux et positifs, cela l'est moins pour Word2Vec, notre modèle de contrôle. Dans ce cas, nous avons fait le choix d'utiliser les k valeurs les plus élevées de la valeur absolue des composantes des vecteurs.

Pour contrôler la qualité de l'espace une fois ce protocole d'éparsification effectué, nous utilisons l'évaluation en similarité, soit la corrélation entre la similarité entre paires de mots dans les espaces appris et la similarité donnée par des humains. Les jeux de données sélectionnés correspondent dans la mesure du possible à une variété de relations : MEN, WS353, SCWS. Pour évaluer la stabilité des vecteurs produits par les deux modèles SPINE et SINr, le second critère d'interprétabilité que nous considérons, nous reproduisons ces expérimentations dix fois et présentons des résultats consolidés.

Puisque les jeux de données en similarité disponibles portent très majoritairement sur l'anglais, le British National Corpus (BNC) et l'Open American National Corpus (OANC) sont choisis comme corpus d'apprentissage. Le BNC compte environ 100 millions de tokens, tandis que la section de données textuelles de OANC en compte environ 11 millions. Les deux corpus sont composites en domaines et en genres textuels. Un prétraitement commun est appliqué aux deux corpus avec la bibliothèque spaCy : tokenisation et regroupement des entités nommées, suppression des mots de longueur inférieure ou égale à deux caractères comme approximation de suppression des mots vides, suppression de la ponctuation et des caractères numériques, et conservation dans le vocabulaire à représenter des items lexicaux apparaissant au moins vingt fois dans les corpus. Après ce prétraitement,

les données considérées sont : 20 814 types pour 4 millions de tokens pour OANC, 58 687 types pour 40 millions de tokens pour BNC.

4 Résultats et discussions

Résultats en stabilité. Les premiers résultats que nous considérons Table 1 ont le double emploi d'établir un point de référence sur les résultats en similarité des modèles avant l'application de notre protocole d'éparsification, et de proposer un indice sur la stabilité des représentations. Dix représentations sont produites avec chaque modèle (sur les mêmes données avec les mêmes hyperparamètres) et sont évaluées en similarité sur les trois jeux de données choisis.

BNC	MEN		WS353		SCWS	
	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ
Word2Vec	$0,72$	2e−3	$0,65$	5e−3	$0,57$	2e−3
SPINE	$0,65$	6e−3	$0,57$	1e−2	$0,60$	4e−3
SINr	$0,66$	6e−4	$0,63$	2e−3	$0,54$	1e−3
OANC	MEN		WS353		SCWS	
	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ
Word2Vec	$0,43$	2e−3	$0,50$	5e−3	$0,46$	3e−3
SPINE	$0,36$	9e−3	$0,43$	1e−2	$0,39$	1e−2
SINr	$0,39$	8e−4	$0,44$	2e−3	$0,39$	2e−3

TABLE 1 – Stabilité des résultats en similarité sur le corpus BNC en haut, sur OANC en bas. Le coefficient de corrélation de Pearson moyen et l'écart-type sont donnés pour 10 exécutions.

Les trois modèles proposent des performances du même ordre de grandeur, quoique légèrement meilleures pour Word2Vec. SPINE semble légèrement plus instable que Word2Vec dont il est tributaire, et que SINr. La variabilité de SINr et de Word2Vec sont comparables. Quoique légère, l'instabilité constatée sur ces différents modèles nuit à leur interprétabilité.

Performance en similarité en fonction de la parcimonie. Les résultats récapitulés dans la Figure 1 permettent d'observer les performances en similarité en fonction du nombre de composantes conservées par vecteur suivant notre protocole d'éparsification. Il est d'abord notable que les trois modèles proposent des performances comparables à celles obtenues en Table 1 en similarité malgré l'éparsification, et ce jusqu'à ne conserver que cinquante dimensions activées par vecteur. Il apparaît même que pour SINr, le processus d'éparsification améliore les performances sur cette tâche, possiblement en retirant du bruit présent dans les représentations originales. La relation entre parcimonie et performance n'est donc pas nécessairement un compromis. De même, la conservation de résultats convenables malgré l'éparsification forcée sur notre modèle de contrôle, bien qu'il soit originellement dense, est un résultat intéressant sur la répartition de l'information sémantique dans celui-ci.

Pour tendre vers l'horizon de parcimonie fixée Section 1, soit dix dimensions, l'expérience est égale-

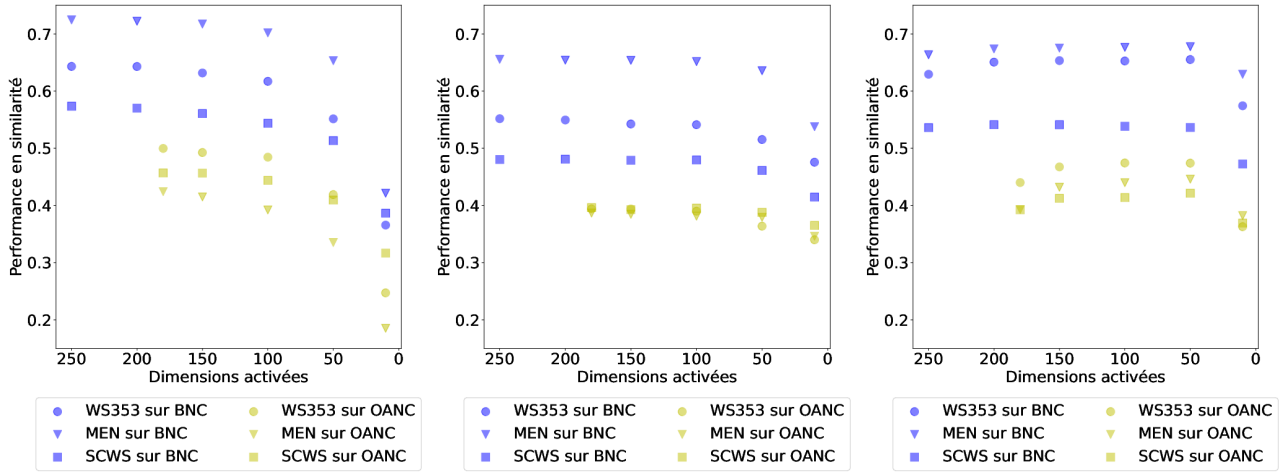

FIGURE 1 – La similarité(en ordonnée, corrélation de Pearson) en fonction du nombre de dimensions conservées par vecteur (en abscisse) pour `Word2Vec` à gauche, `SPINE` au milieu et `SINr` à droite. En jaune, les performances sur le corpus OANC, en bleu celles sur le corpus BNC.

ment menée à ce palier. Bien que les performances baissent significativement pour trois modèles, et particulièrement pour `Word2Vec`, une partie importante de l'information sémantique semble conservée dans ces dix dimensions dans la mesure où elles permettent de résoudre, au moins partiellement, la tâche de similarité. Quoique cette restriction de dix dimensions activées sur les vecteurs ne semble pas intéressante pour des utilisations en aval des vecteurs (considérant la chute de performance), elle permet de construire des vecteurs interprétables. Par ailleurs, cette très faible quantité de composantes rend plus plausible la compatibilité de ces représentations aux modèles théoriques utilisant des traits sémantiques, ouvrant ainsi d'éventuelles opportunités empiriques.

5 Conclusion

Dans cet article, nous proposons de définir l'interprétabilité au niveau des vecteurs, et plus seulement au niveau des dimensions de l'espace de plongement. Nous proposons ainsi un protocole d'évaluation basé sur deux critères : la stabilité des représentations produites, et la contrainte de parcimonie que nous redéfinissons en accord avec la plausibilité psycholinguistique. Il nous paraît en effet souhaitable, non seulement de pouvoir trouver une cohérence sémantique interne à une dimension de description dans une représentation du lexique, mais de pouvoir décrire chaque mot avec un faible nombre de ces dimensions. Nous faisons l'hypothèse que les vecteurs correspondant à ces contraintes sont interprétables par un locuteur, puisqu'il devient possible de manipuler ce faible nombre de dimensions en mémoire de travail.

Il apparaît que les modèles de plongement lexicaux interprétables conservent des résultats intéressants sur la tâche de similarité même en forçant des parcimonies élevées, la méthode `SINr` bénéficiant même du processus d'éparsification appliqué. Ce résultat permet de reconsidérer le compromis entre interprétabilité et performance pour les représentations lexicales. La construction de représentations lexicales aux vecteurs interprétables ouvre par ailleurs la perspective de mettre en relation les modèles théoriques décrivant le lexique depuis des traits sémantiques avec des plongements lexicaux.

Remerciements

Ces travaux ont été financés dans le cadre du projet ANR-21-CE23-0010 DIGING.

Références

BLONDEL V. D., GUILLAUME J. L., LAMBIOTTE R. & LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, **2008**(10), P10008.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DIGAN W., NÉVÉOL A., NEURAZ A., WACK M., BAUDOIN D., BURGUN A. & RANCE B. (2020). Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, **28**(3), 504–515. DOI : 10.1093/jamia/ocaa261.

FARUQUI M., TSVETKOV Y., YOGATAMA D., DYER C. & SMITH N. (2015). Sparse overcomplete word vector representations. *arXiv preprint arXiv :1506.02004*.

GARRARD P., LAMBON RALPH M., HODGES J. & PATTERSON K. (2001). Prototypicality, distinctiveness, and intercorrelation : Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, **18**(2), 125–174.

HELLRICH J. & HAHN U. (2016a). An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In *SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, p. 111–117.

HELLRICH J. & HAHN U. (2016b). Bad Company Neighborhoods in neural embedding spaces considered harmful. In *Conference on Computational Linguistics*, p. 2785–2796.

JACKENDOFF R. (1983). *Semantic and Cognition*. MIT Press.

LEE D. D. & SEUNG H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, **401**, 788–791.

LEVY O. & GOLDBERG Y. (2014). Neural word embedding as implicit matrix factorization. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. LAWRENCE & K. WEINBERGER, Éds., *Advances in Neural Information Processing Systems*, volume 27 : Curran Associates, Inc.

MCRAE K., CREE G. S., SEIDENBERG M. S. & MCNORGAN C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, **37**(4), 547.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MILLER G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review*, **63**(2), 81–97.

MURPHY B., TALUKDAR P. & MITCHELL T. (2012). Learning effective and interpretable semantic models using non-negative sparse embedding. In *Conference on Computational Linguistics*, p. 1933–1950.

PALMER S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, **9**(4), 441–474. DOI : https ://doi.org/10.1016/0010-0285(77)90016-0.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, p. 1532–1543.

PETERSON L. & PETERSON M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, **58**(3), 193. DOI : 10.1037/h0049234.

PIERREJEAN B. (2020). *Qualitative Evaluation of Word Embeddings : Investigating the Instability in Neural-Based Models*. Thèse de doctorat, Université Toulouse 2 - Jean Jaurès.

POTTIER B. (1963). *Recherches sur l'analyse sémantique en linguistique et en traduction mécanique*. Publications linguistiques de la Faculté des lettres et sciences humaines de Nancy.

PROUTEAU T., CONNES V., DUGUÉ N., PEREZ A., LAMIREL J.-C., CAMELIN N. & MEIGNIER S. (2021). SINr : Fast Computing of Sparse Interpretable Node Representations is not a Sin ! In *Intelligent Data Analysis*, volume 12695, p. 325–337.

PROUTEAU T., DUGUÉ N., CAMELIN N. & MEIGNIER S. (2022). Are embedding spaces interpretable ? results of an intrusion detection evaluation on a large french corpus. In *Language Ressources and Evaluation Conference*.

RASTIER F. (2009). Principes et conditions de la sémantique componentielle. In *Sémantique interprétative*, Formes sémiotiques, p. 17–37.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2021). A Primer in BERTology : What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866.

RUDIN C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, **1**(5), 206–215.

SENEL L. K., UTLU I., YUCESOY V., KO.C A. & CUKUR T. (2018). Semantic structure and interpretability of word embeddings. *Transactions on Audio, Speech, and Language Processing*, **26**(10), 1769–1779.

SHIN J., MADOTTO A. & FUNG P. (2018). Interpreting word embeddings with eigenvector analysis. *Advances in Neural Information Processing Systems*, **32**.

SUBRAMANIAN A., PRUTHI D., JHAMTANI H., BERG-KIRKPATRICK T. & HOVY E. (2018). Spine : Sparse interpretable neural embeddings. In *AAAI conference on artificial intelligence*, volume 32.

ZAFAR M. R. & KHAN N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, **3**(3), 525–541.