État de l'art du changement sémantique à partir de plongements contextualisés

Syrielle Montariol* — **Antoine Doucet** ** — **Alexandre Allauzen** ***

* *LISN-CNRS, Université Paris-Saclay*
** *La Rochelle Université*
*** *ESPCI, Dauphine Université - PSL*

RÉSUMÉ. *Les changements lexico-sémantiques — des variations temporelles dans l'usage et la signification des mots — reflètent l'évolution de divers aspects de la société tels que l'environnement technologique et culturel. Détecter et comprendre ces changements est utile, par exemple, en lexicographie et en sociolinguistique. Ce domaine d'étude a rapidement évolué avec l'essor de la sémantique distributionnelle et a connu un élan d'intérêt au cours des dernières années, avec l'usage des plongements neuronaux. Plus récemment, les modèles de langue pré-entraînés et les plongements contextualisés élargissent le champs des possibles. Cet article résume l'état de l'art sur la détection et l'analyse de ces phénomènes en se concentrant sur les plongements contextualisés. En particulier, nous abordons l'aspect pratique de ces systèmes à travers le passage à l'échelle des différentes méthodes, en vue de les appliquer à de grands corpus ou de larges vocabulaires.*

ABSTRACT. *Lexico-semantic changes — temporal variations in the use and meaning of words — reflect the evolution of various aspects of society such as the technological and cultural environment. Detecting and understanding these changes is useful, for example, in lexicography and sociolinguistics. This field of study has evolved rapidly with the rise of distributional semantics and has experienced a surge of interest in recent years with the use of neural embeddings. More recently, pre-trained language models and contextualised embeddings are broadening the field of possibilities. This paper summarizes the state of the art on the detection and analysis of these phenomena by focusing on contextualized embeddings. In particular, we address the practical aspect of these systems through the scaling of the different methods in order to apply them to large corpora or large vocabularies.*

MOTS-CLÉS : *Changement sémantique, Plongements contextualisés, Diachronie.*

KEYWORDS: *Semantic change, Contextualized embeddings, Diachrony.*

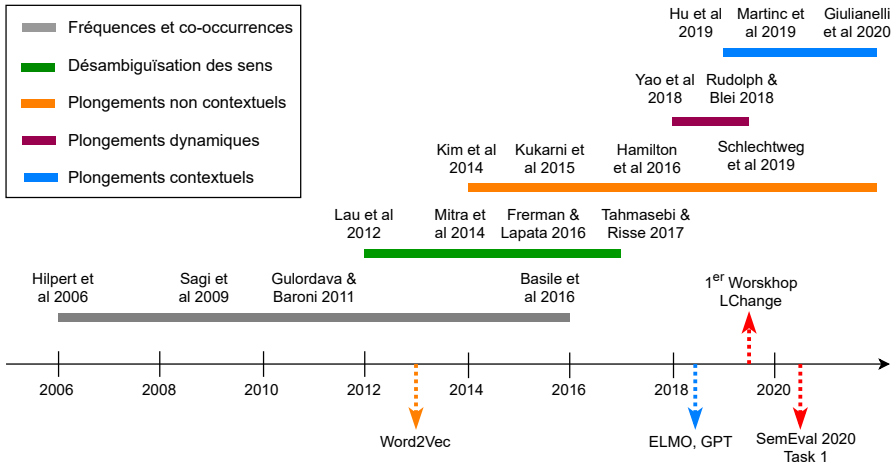

Figure 1. *Chronologie des méthodes de détection du changement sémantique.*

1. Introduction

L'évolution de la signification des mots à travers le temps est appelée *diachronie*. Ce domaine a connu un regain d'intérêt ces dernières années, avec la publication consécutive de trois articles de revue de littérature (Tahmasebi *et al.*, 2018 ; Kutuzov *et al.*, 2018 ; Tang, 2018). Puis avec l'organisation du premier workshop sur l'évolution du langage (LChange 2019) et d'une tâche d'évaluation dédiée (SemEval 2020 Task 1 (Schlechtweg *et al.*, 2020)). Une vue d'ensemble de la chronologie du domaine du changement sémantique se trouve sur la Figure 1. Pour cette tâche, nous considérons un corpus de documents, chacun étant associé à une date. Nous le divisons en plusieurs périodes de temps, selon la granularité souhaitée. Pour un mot cible, nous extrayons un signal — l'information sémantique — à partir de tous les contextes dans lesquels il est utilisé dans chaque période. Le changement sémantique est l'évolution de ce signal dans le temps. Les méthodes présentées dans les sections suivantes visent à extraire ce signal, dans le but d'identifier les mots changeant de sens au cours du temps dans un corpus.

Les premières approches pour la modélisation diachronique reposent sur l'évolution des fréquences et co-occurrences de mots (Gulordava et Baroni, 2011). Suite à la généralisation des plongements neuronaux, des modèles de plongements lexicaux *diachroniques* ont émergé. La méthode la plus simple consiste à apprendre une matrice de plongements sur la première période d'un corpus temporel, et de l'affiner *incrémentalement* à chaque période (Kim *et al.*, 2014). Une autre méthode consiste à apprendre une matrice de plongements de mots *indépendamment* sur chaque période, puis aligner les espaces de représentation pour rendre les plongements comparables (Hamilton *et al.*, 2016 ; Kulkarni *et al.*, 2015 ; Schlechtweg *et al.*, 2019). Enfin, les modèles de plongements *dynamiques* contrôlent l'évolution des plonge-

ments dans le temps avec un processus de diffusion (Rudolph et Blei, 2018 ; Bamler et Mandt, 2017 ; Jawahar et Seddah, 2019 ; Yao *et al.*, 2018).

Ces méthodes présentent différentes limites, notamment une grande sensibilité au bruit (Shoemark *et al.*, 2019 ; Kaiser *et al.*, 2020) et aux variations de fréquence (Dubossarsky *et al.*, 2017). Plus important encore, elles résument tous les usages d'un mot dans un vecteur unique à chaque période, sans tenir compte de la possibilité d'avoir plusieurs sens distincts dans le corpus. Des méthodes de désambiguïsation du sens des mots peuvent être appliquées pour pallier à ce problème (Mitra *et al.*, 2015 ; Tahmasebi et Risse, 2017 ; Frermann et Lapata, 2016). Plus récemment, les plongements contextualisés générés à partir de modèles de langue pré-entraînés tels que BERT (Devlin *et al.*, 2019) donnent une nouvelle perspective à ce problème. Par le présent article, nous complétons les précédentes revues de l'état de l'art en présentant les méthodes utilisées dans la littérature pour quantifier le degré de changement sémantique d'un mot dans un corpus à l'aide de plongements contextualisés.

2. Plongements contextualisés et changement sémantique

Les modèles de langue, pré-entraînés sur de grandes quantités de texte, génèrent des plongements tenant compte du contexte dans lequel le mot apparaît. Ils permettent une amélioration de la précision sur de nombreuses tâches de TAL.

2.1. *Modèles de langue pré-entraînés et extraction des plongements*

Le premier modèle de langue pré-entraîné à avoir été utilisé pour le changement sémantique (Hu *et al.*, 2019), et aussi le plus communément utilisé, est BERT (Devlin *et al.*, 2019). It possède une architecture multicouche de Transformers bidirectionnels (Vaswani *et al.*, 2017). Le modèle ELMo (Peters *et al.*, 2018) est aussi employé pour cette tâche (Kutuzov et Giulianelli, 2020 ; Karnysheva et Schwarz, 2020 ; Rodina *et al.*, 2020). Son architecture plus légère — un LSTM bidirectionnel à deux couches — et son nombre de paramètres bien inférieur permettent de l'entraîner en totalité sur les corpus étudiés, ce qui en fait un modèle de choix pour les corpus peu standard (e.g. lemmatisés ou portant sur un domaine spécifique). Pour finir, XLM-R (Conneau *et al.*, 2020) est parfois utilisé lorsque les corpus analysés sont dans une langue autre que l'anglais (Arefyev et Zhikov, 2020 ; Rother *et al.*, 2020 ; Cuba Gyllensten *et al.*, 2020).

Afin d'obtenir des plongements contextualisés, les documents du corpus sont divisés en séquences, tokenisés, et donnés en entrée du modèle de langue choisi. Une séquence de plongements est générée à partir des couches de sortie du modèle. Pour BERT et XLM-R, ce sont le plus souvent les 4 dernières couches qui sont sommées ou concaténées pour obtenir les plongements. En effet, l'information sémantique est plutôt capturée dans les couches supérieures des modèles de langue (Devlin *et al.*, 2019 ; Jawahar *et al.*, 2019). Alternativement, il est possible d'utiliser unique-

ment la dernière couche (Rother *et al.*, 2020). Kutuzov (2020) effectue une brève analyse du choix des couches à utiliser (toutes, les 4 dernières, seulement la dernière); il montre que la dernière couche permet en général d'obtenir les meilleurs scores de détection de changement sémantique pour les modèles ELMo et BERT.

2.2. *Mesure du changement sémantique*

Suite à l'apparition des modèles de langue présentés précédemment, le premier article les appliquant au changement sémantique se base sur l'identification des différents sens d'un mot dans un corpus temporel (Hu *et al.*, 2019). Les plongements contextualisés sont utilisés de manière supervisée : pour un mot-cible polysémique, une représentation pour chaque sens est générée en appliquant un modèle BERT pré-entraîné à des phrases comportant le mot utilisé selon le sens en question. Ce modèle est ensuite appliqué à un corpus temporel; chaque occurrence du mot-cible est associé avec le sens dont le plongement est le plus proche. Enfin, la proportion de chaque sens est calculée dans les strates temporelles successives, révélant l'évolution de la distribution des sens pour le mot cible. Cette méthode exige que l'ensemble des sens de chaque mot cible soit connu à l'avance, et ne peut être appliquée qu'à un nombre réduit de mots.

Par la suite, des méthodes non supervisées ont été utilisées. Nous en exposons 4 dans les paragraphes suivants, et les illustrons dans les figures 2 à 5. Pour toutes ces méthodes, l'objectif est de comparer les plongements contextualisés d'un mot-cible extraits de deux strates temporelles d'un corpus.

Distance moyenne par paire (DMP, Figure 2). Une première méthode consiste à calculer cette distance entre tous les plongements contextualisées d'un mot dans deux périodes (Giulianelli *et al.*, 2020). On note $E_w^{(t)}$ la matrice des plongements du mot w au temps t, avec $N^{(t)}$ sa dimension (le nombre d'occurrences du mot dans la strate t):

$$\mathrm{DMP}(E_w^{(t_1)}, E_w^{(t_2)}) = \frac{1}{N^{(t_1)} N^{(t_2)}} \sum_{\substack{u_i \in E_w^{(t_1)} \\ u_j \in E_w^{(t_2)}}} d(u_i, u_j) \qquad [1]$$

La mesure d la plus utilisée est la distance cosinus. Plus rarement, on trouve la distance de Canberra (Lance et Williams, 1967) utilisée par Giulianelli *et al.* (2020), et la distance euclidienne utilisée par Pömsl et Lyapin (2020).

La DMP permet de quantifier la variation contextuelle d'un mot en comparant toutes les occurrences d'un mot dans une période à toutes celles d'une autre période. À ce titre, cette méthode est fortement liée à la diversité intra-période; les mots dont le contexte est très diversifié, comme les mots-outils, peuvent présenter une DMP particulièrement élevée sans pour autant avoir subi de changement sémantique.

Distance entre les moyennes (MD, Figure 3). Cette méthode consiste à calculer la moyenne des plongements contextualisés d'un mot dans chaque strate temporelle (Martinc *et al.*, 2020b). On obtient une unique représentation vectorielle pour chaque période; elles sont le plus souvent comparées avec la distance cosinus.

$$\mathrm{MD}(E_w^{(t_1)}, E_w^{(t_2)}) = d\left(\frac{\sum_{u_i \in E_w^{(t_1)}} u_i}{N^{(t_1)}}, \frac{\sum_{u_j \in E_w^{(t_2)}} u_j}{N^{(t_2)}}\right) \quad [2]$$

Pour mesurer la similarité entre les plongements, on utilise en général l'opposé de la distance $(1 - d)$ et parfois l'inverse $(\frac{1}{d})$ (Kutuzov et Giulianelli, 2020).

La DM revient à générer des plongements diachroniques non contextualisés. En conséquence, toute information sur la diversité contextuelle intra-cluster est perdue. En contrepartie, cette méthode est moins sensible au bruit que la précédente.

Clustering (Figure 4). Le but de cette méthode est de résumer l'information contenue dans les plongements contextuels d'un mot dans une période en regroupant ces plongements en clusters (Giulianelli *et al.*, 2020). La méthode la plus courante consiste à appliquer le clustering sur les plongements extraits de toutes les strates temporelles conjointement, afin de comparer la distribution des clusters entre les différentes strates.

Les algorithmes utilisés dans la littérature sont k-means (Giulianelli *et al.*, 2020), propagation par affinité (Martinc *et al.*, 2020a), DBSCAN (Karnysheva et Schwarz, 2020), clustering agglomératif (Arefyev et Zhikov, 2020), HDBSCAN et GMM (Rother *et al.*, 2020). En pratique, k-means est utilisé le plus couramment; il bénéficie d'une faible complexité, ce qui est adéquat pour regrouper un grand nombre de plongements (lorsque le mot cible est fréquent) en grande dimension.[1] Le nombre de clusters est en général choisi à l'aide du score de silhouette (Rousseeuw, 1987). Il peut aussi être fixé arbitrairement (e.g. 8 pour Cuba Gyllensten *et al.* (2020)). À l'inverse, les algorithmes tels que la propagation par affinité (Frey et Dueck, 2007) et DBSCAN déduisent automatiquement le nombre de clusters le plus adapté. Ils mènent généralement à un très grande nombre de clusters dont les tailles sont inégalement distribuées. Une étape supplémentaire peut être ajoutée après le clustering, consistant à supprimer les clusters minoritaires et éloignés, et fusionner les clusters les plus similaires (Martinc *et al.*, 2020a); cela revient à diminuer le bruit en se concentrant sur les usages "principaux" et améliore la précision du clustering pour la tâche.

Suite au clustering, on déduit la distribution de chaque cluster C dans chaque période $\mathcal{T}$, que l'on note $p(C|\mathcal{T}, W)$, à partir du nombre de plongements dans chaque cluster et pour chaque période normalisé par le nombre total d'occurrences du mot dans le corpus. Ces distributions sont comparées entre deux périodes en utilisant la divergence de Jensen-Shannon (JSD, (Lin, 1991)), une mesure de dissimilarité entre deux distributions de probabilité (Giulianelli *et al.*, 2020). Ainsi, cette méthode capte

1. Par exemple, la dimension des plongements de BERT est de 768.

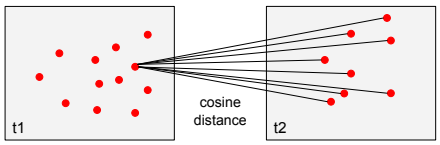

Figure 2. *Distance moyenne par paire*

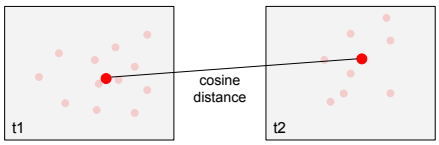

Figure 3. *Distance entre les moyennes*

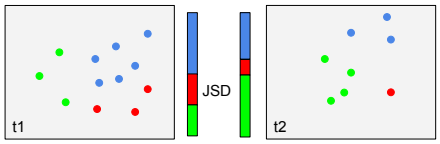

Figure 4. *Clustering*

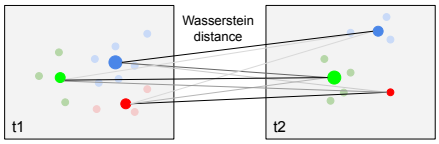

Figure 5. *Transport Optimal*

la variabilité du contexte d'un mot en décomposant ses représentations en une distribution de faible dimension; cependant, l'information sémantique apprise par le modèle et enregistrée dans les plongements est perdue, étant donné que seule la différence entre les distributions des clusters est quantifiée.

Il est possible d'effectuer un clustering séparé pour chaque strate temporelle du corpus, puis d'aligner les clusters entre les strates (Kanjirangat *et al.*, 2020). La distance entre deux cluster est mesurée par la distance euclidienne entre leurs centroïdes, et chaque cluster d'une strate est associé à celui qui lui est le plus similaire dans l'autre strate. L'alignement optimal entre les clusters est obtenu en suivant l'algorithme Hongrois (Kuhn, 1955). En pratique, cette méthode aboutit à des résultats très similaires à ceux obtenus à partir d'un clustering unique.

Transport optimal (Figure 5). Montariol *et al.* (2021) proposent une méthode située à l'intersection de la moyenne et du clustering, qui permet de prendre en compte la position des clusters dans l'espace lors de la mesure de la distance, en s'appuyant sur le cadre du transport optimal. L'ensemble des plongements contextualisés d'un mot sont regroupés via un clustering. Ensuite, les auteurs calculent la moyenne de tous les plongements d'une période à l'intérieur d'un cluster, formant des *centroïdes*, pondérés par le nombre de plongements dans le cluster associé. Ces centroïdes pondérés sont comparés entre les périodes à l'aide de la distance de Wasserstein (Solomon, 2018); l'idée est de déterminer le plan de transport qui minimise l'effort requis pour transporter la masse des centroïdes pondérés de la période 1 sur ceux de la période 2. Notons que puisque ce sont deux ensembles pondérés de centroïdes au lieu de deux distributions qui sont comparées, les clusters des différentes périodes n'ont pas besoin d'être alignés.

3. Passage à l'échelle

La principale limite des méthodes de DMP, de clustering et de transport optimal présentées dans la section précédente est le passage à l'échelle en termes de mémoire et de temps de calcul. En effet, le clustering doit être appliquée pour chaque mot du corpus séparément et les plongements de toutes les occurrences du mot doivent être stockés en mémoire. Pour les grands corpus à large vocabulaire, où certains mots peuvent apparaître des millions de fois, la faisabilité de ces méthodes se trouve fortement limitée. La méthode de moyenne des plongements n'est pas confrontée à cette limitation, car les plongements de chaque occurrence de mot ne sont pas rassemblés dans une liste, mais plutôt sommées au fur et à mesure de l'extraction. Cependant, elle implique de renoncer à l'information fournie par la diversité contextuelle intra-période, qui peut être précieuse pour la précision de la méthode et pour l'interprétation. En pratique, ces limites de complexité et de mémoire réduisent considérablement les applications possibles des méthodes dans un cadre exploratoire, pour des tâches telles que l'identification des mots ayant le plus fort degré de changement parmi l'ensemble du vocabulaire ou la mesure du degré de changement sémantique des mots très fréquents.

Sélection préliminaire de mot-cibles. Une solution simple pour faire face au problème de passage à l'échelle consiste à ajouter une étape préliminaire qui serait appliquée sur le vocabulaire complet d'un corpus, pour identifier les mots qui ont potentiellement subi un changement sémantique (Martinc *et al.*, 2020a). Au cours de cette étape préliminaire, on peut effectuer la méthode de calcul de la moyenne (MD), qui ne subit pas de problème de passage à l'échelle. La définition d'un seuil (par exemple, sous la forme d'une fraction de la taille du vocabulaire complet) permet d'obtenir une liste réduite de mots cibles, à partir de la liste de tous les mots classés par degré de changement sémantique. Ainsi, le vocabulaire est filtré avant d'appliquer des méthodes plus lourdes et plus précises telles que le clustering.

Extraction d'un nombre réduit de plongements. Au lieu de garder en mémoire autant de vecteurs que d'occurrences d'un mot, Montariol *et al.* (2021) proposent une méthode de regroupement-moyenne en ne stockant qu'un nombre limité N de plongements pour chaque strate dans une liste L. À chaque nouvelle occurrence du mot dans la strate, son plongement e_{new} est additionné au vecteur e_m qui lui est le plus similaire dans L.[2] Le nombre d'éléments additionnés dans e_m est incrémenté pour normaliser chaque élément de la liste à la fin de l'extraction. Les auteurs utilisent un seuil $N = 200$, qui offre un compromis raisonnable entre mémoire et performance.

Réduction de dimension. Dans le but d'utiliser des algorithmes de clustering complexes tels que Hierarchical Density-Based Clustering (HDBSCAN), qui ne sont pas adaptés à des données en grande dimension (comme ceux issus de BERT), Rother *et al.* (2020) appliquent une méthode de réduction de dimension après l'extraction des plongements contextualisés. Pour cela, ils utilisent d'abord un auto-encodeur pour réduire la dimension des plongements à 20, puis l'algorithme UMAP (similaire à t-

2. En utilisant la distance cosinus: $e_m = \arg\max_{e_i \in L} \cos(e_i, e_{\text{new}})$.

SNE, avec la distance cosinus comme mesure entre les plongements) (McInnes *et al.*, 2018) pour atteindre une dimension de 10. Les auteurs montrent que la réduction de la dimension contribue à la stabilité du système dans son ensemble.

4. Conclusion

La campagne d'évaluation SemEval 2020 Task 1, qui porte sur la détection de changement sémantique de façon non supervisée, a montré qu'en moyenne sur plusieurs corpus de test, les plongements non contextuels surpassent les méthodes basées sur les plongements contextualisés. Cette constatation vient s'ajouter au fait que les modèles de langue pré-entraînés sont très lourds comparés aux modèles de plongements non contextuels. Néanmoins, lorsque le modèle de langue, les hyperparamètres et la méthode de détection sont soigneusement choisis pour chaque corpus de test, les plongements contextualisés peuvent surpasser leurs équivalents non contextuels (Kutuzov et Giulianelli, 2020). De plus, un ensemble de ces deux types de méthodes peut améliorer encore les résultats (Pömsl et Lyapin, 2020). En outre, les différentes méthodes rapportées dans cet article sont aptes à aborder différents aspects de la tâche de détection du changement sémantique; certaines se concentrent sur le degré de variation contextuelle (distance moyenne par paire), d'autres sur l'apparition ou la disparition d'un usage spécifique du mot (clustering). Ce dernier type de méthode permet une interprétation plus fine du changement sémantique, en identifiant le sens concerné par le changement. Néanmoins, il est important de noter que les clusters obtenus à partir des représentations d'un mot ne reflètent pas naturellement les différents sens du mot d'un point de vue lexicographique, mais plutôt les différentes façons dont il est utilisé. En effet, les modèles de langue pré-entraînés ne retiennent pas seulement les informations sémantiques pour les représentations contextualisées ; ils sont entre autres fortement influencés par la syntaxe (Reif *et al.*, 2019). Le terme de *changement sémantique* est donc en général utilisé, dans la littérature utilisant des plongements contextualisés, pour désigner des *variation contextuelles*.

De façon générale lors de l'évaluation des méthodes de détection de changement sémantique, on observe que la meilleure méthode diffère selon le corpus de test (Schlechtweg *et al.*, 2020). Pour mieux comprendre en quoi les méthodes diffèrent, un contrôle précis de l'évaluation est nécessaire. En particulier, il serait intéressant d'évaluer les différentes méthodes sur diverses sous-tâches de changement sémantique (mesure de la vitesse du changement sémantique, identification du type de changement sémantique…). L'étude de l'impact de la polysémie et de la fréquence, et de la sensibilité au bruit en général, est également cruciale. Or, les corpus annotés en changement sémantiques sont rares et limités, ne permettant pas de couvrir les différentes tâches ni d'étudier de façon approfondie l'impact des éléments tels que le degré de polysémie des mots. Afin de contrôler tous les aspects du changement sémantiques mentionnés, une solution à explorer consiste à générer un corpus comprenant des changements sémantiques synthétiques (Shoemark *et al.*, 2019), construisant ainsi un cadre d'évaluation étendu en complément des corpus annotés existants.

5. Bibliographie

Arefyev N., Zhikov V., "BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change Detection", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 171-179, December, 2020.

Bamler R., Mandt S., "Dynamic Word Embeddings", *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, PMLR, International Convention Centre, Sydney, Australia, p. 380-389, Aug, 2017.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V., "Unsupervised cross-lingual representation learning at scale", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440-8451, July, 2020.

Cuba Gyllensten A., Gogoulou E., Ekgren A., Sahlgren M., "SenseCluster at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 112-118, December, 2020.

Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: pre-training of deep bidirectional transformers for language understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, p. 4171-4186, June, 2019.

Dubossarsky H., Weinshall D., Grossman E., "Outta control: laws of semantic change and inherent biases in word representation models", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1136-1145, 2017.

Frermann L., Lapata M., "A Bayesian model of diachronic meaning change", *Transactions of the Association for Computational Linguistics*, vol. 4, p. 31-45, 2016.

Frey B. J., Dueck D., "Clustering by passing messages between data points", *Science*, vol. 315, n^{o} 5814, p. 972-976, 2007.

Giulianelli M., Del Tredici M., Fernández R., "Analysing lexical semantic change with contextualised word representations", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3960-3973, July, 2020.

Gulordava K., Baroni M., "A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus.", *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, p. 67-71, 2011.

Hamilton W. L., Leskovec J., Jurafsky D., "Diachronic word embeddings reveal statistical laws of semantic change", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1489-1501, 2016.

Hu R., Li S., Liang S., "Diachronic sense modeling with deep contextualized word embeddings: an ecological view", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, p. 3899-3908, July, 2019.

Jawahar G., Sagot B., Seddah D., "What does BERT learn about the structure of language?", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, p. 3651-3657, July, 2019.

Jawahar G., Seddah D., "Contextualized diachronic word representations", p. 35-47, 01, 2019.

Kaiser J., Schlechtweg D., Papay S., Schulte im Walde S., "IMS at SemEval-2020 Task 1: How Low Can You Go? Dimensionality in Lexical Semantic Change Detection", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 81-89, December, 2020.

Kanjirangat V., Mitrovic S., Antonucci A., Rinaldi F., "SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 214-221, December, 2020.

Karnysheva A., Schwarz P., "TUE at SemEval-2020 Task 1: Detecting Semantic Change by Clustering Contextual Word Embeddings", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 232-238, December, 2020.

Kim Y., Chiu Y.-I., Hanaki K., Hegde D., Petrov S., "Temporal analysis of language through neural language models", *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, p. 61-65, 2014.

Kuhn H. W., "The Hungarian method for the assignment problem", *Naval research logistics quarterly*, vol. 2, n° 1-2, p. 83-97, 1955.

Kulkarni V., Al-Rfou R., Perozzi B., Skiena S., "Statistically significant detection of linguistic change", *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, International World Wide Web Conferences Steering Committee, p. 625-635, 2015.

Kutuzov A., "Distributional word embeddings in modeling diachronic semantic change", *PhD Thesis*, 2020.

Kutuzov A., Giulianelli M., "UiO-UvA at SemEval-2020 Task 1: Contextualised embeddings for lexical semantic change detection", *ArXiv*, 2020.

Kutuzov A., Øvrelid L., Szymanski T., Velldal E., "Diachronic word embeddings and semantic shifts: a survey", *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1384-1397, 2018.

Lance G. N., Williams W. T., "Mixed-data classificatory programs I: Agglomerative Systems", *Australian Computer Journal*p. 15-20, 1967.

Lin J., "Divergence Measures Based on the Shannon Entropy", *IEEE Trans. Inf. Theor.*, vol. 37, n° 1, p. 145–151, September, 1991.

Martinc M., Montariol S., Zosa E., Pivovarova L., "Capturing evolution in word usage: just add more clusters?", *Companion Proceedings of the Web Conference 2020*, WWW '20, Association for Computing Machinery, New York, NY, USA, p. 343–349, 2020a.

Martinc M., Novak P. K., Pollak S., "Leveraging contextual embeddings for detecting diachronic semantic shift", *LREC*, 2020b.

McInnes L., Healy J., Melville J., "Umap: Uniform manifold approximation and projection for dimension reduction", *arXiv preprint arXiv:1802.03426*, 2018.

Mitra S., Mitra R., Maity S. K., Riedl M., Biemann C., Goyal P., Mukherjee A., "An automatic approach to identify word sense changes in text media across timescales", *Natural Language Engineering*, vol. 21, p. 773-798, 2015.

Montariol S., Martinc M., Pivovarova L., "Scalable and Interpretable Semantic Change Detection", *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June, 2021.

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L., "Deep contextualized word representations", *Proceedings of NAACL-HLT*, p. 2227-2237, 2018.

Pömsl M., Lyapin R., "CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 180-186, December, 2020.

Reif E., Yuan A., Wattenberg M., Viegas F. B., Coenen A., Pearce A., Kim B., "Visualizing and measuring the geometry of BERT", *Advances in Neural Information Processing Systems 32*, p. 8594-8603, 2019.

Rodina J., Trofimova Y., Kutuzov A., Artemova E., "ELMo and BERT in semantic change detection for Russian", 2020.

Rother D., Haider T., Eger S., "CMCE at SemEval-2020 Task 1: Clustering on Manifolds of Contextualized Embeddings to Detect Historical Meaning Shifts", *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 187-193, December, 2020.

Rousseeuw P. J., "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of computational and applied mathematics*, vol. 20, p. 53-65, 1987.

Rudolph M., Blei D., "Dynamic embeddings for language evolution", *Proceedings of the 2018 World Wide Web Conference*, WWW '18, p. 1003–1011, 2018.

Schlechtweg D., Hätty A., Del Tredici M., Schulte im Walde S., "A wind of change: detecting and evaluating lexical semantic change across times and domains", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, p. 732-746, July, 2019.

Schlechtweg D., McGillivray B., Hengchen S., Dubossarsky H., Tahmasebi N., "SemEval-2020 Task 1: unsupervised lexical semantic change detection", *Proceedings of the 14th International Workshop on Semantic Evaluation*, 2020.

Shoemark P., Liza F. F., Nguyen D., Hale S., McGillivray B., "Room to glo: a systematic comparison of semantic change detection approaches with word embeddings", *Proceedings of EMNLP-IJCNLP 2019*, Hong Kong, China, p. 66-76, November, 2019.

Solomon J., "Optimal transport on discrete domains", *AMS Short Course on Discrete Differential Geometry*, 2018.

Tahmasebi N., Borin L., Jatowt A., "Survey of computational approaches to diachronic conceptual change", *ArXiv*, 2018.

Tahmasebi N., Risse T., "Finding individual word sense changes and their delay in appearance", *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, p. 741-749, 2017.

Tang X., "A state-of-the-art of semantic change computation", *Natural Language Engineering*, vol. 24, n° 5, p. 649–676, 2018.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., "Attention is all you need", *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, 2017.

Yao Z., Sun Y., Ding W., Rao N., Xiong H., "Dynamic word embeddings for evolving semantic discovery", *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, ACM, p. 673-681, 2018.