# Un mot, deux facettes : traces des opinions dans les représentations contextualisées des mots

Aina Garí Soler Matthieu Labeau Chloé Clavel
LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{aina.garisoler,matthieu.labeau,chloe.clavel}@telecom-paris.fr

RÉSUMÉ

La façon dont nous utilisons les mots est influencée par notre opinion. Nous cherchons à savoir si cela se reflète dans les plongements de mots contextualisés. Par exemple, la représentation d' « animal » est-elle différente pour les gens qui voudraient abolir les zoos et ceux qui ne le voudraient pas ? Nous explorons cette question du point de vue du changement sémantique des mots. Nos expériences avec des représentations dérivées d'ensembles de données annotés avec les points de vue révèlent des différences minimes, mais significatives, entre postures opposées [1].

ABSTRACT

**One Word, Two Sides : Traces of Stance in Contextualized Word Representations**

The way we use words is influenced by our opinion. We investigate whether this is reflected in contextualized word embeddings. For example, is the representation of "animal" different between people who would abolish zoos and those who would not ? We explore this question from a Lexical Semantic Change standpoint. Our experiments with BERT embeddings derived from datasets with stance annotations reveal small but significant differences in word representations between opposing stances.

MOTS-CLÉS : Représentations contextualisées, changement sémantique, détection de point de vue.

KEYWORDS: Contextualized representations, semantic change, stance detection.

## 1 Introduction

Nos opinions se reflètent dans notre façon de parler. Les personnes ayant des positions opposées sur un sujet particulier peuvent utiliser des mots différents pour en discuter. Par exemple, seules les personnes contre l'utilisation de masques pendant la pandémie de COVID-19 sont susceptibles de les appeler des « muselières ». Dans cet article, cependant, nous n'étudions pas *quels* mots sont utilisés de part et d'autre : nous comparons plutôt la façon dont les locuteurs qui sont en désaccord sur un sujet utilisent les *mêmes* mots. Plus précisément, nous cherchons à savoir si les modèles contextuels capturent une différence entre la représentation d'un mot (par exemple, « masque ») lorsqu'il est utilisé par des personnes qui sont pour ou contre une certaine cible (par exemple, l'utilisation obligatoire de masques).

Nous abordons cette question du point de vue du changement lexico-sémantique (CLS). Les travaux sur le CLS visent généralement à détecter les changements de sens des mots sur deux périodes de

1. Cet article est une adaptation et traduction de Garí Soler *et al.* (2022).

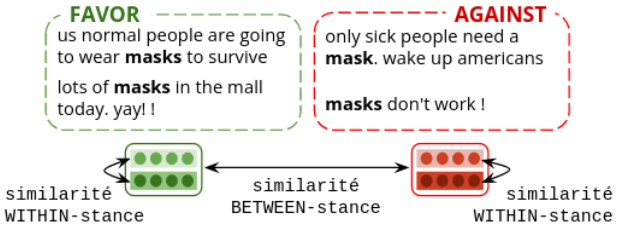


FIGURE 1 – Exemples d'instances de « mask » de l'ensemble de données de position Covid19 (Glandt *et al.*, 2021). Nous comparons la similarité d'utilisation interne aux, et entre, les positions.

temps ou plus (Tahmasebi *et al.*, 2021), mais leurs techniques ont également été employées pour identifier les différences synchroniques dans l'utilisation des mots, par exemple à travers différents âges, sexes, professions (Gonen *et al.*, 2020), domaines (Yin *et al.*, 2018; Schlechtweg *et al.*, 2019) ou cultures (Garimella *et al.*, 2016). Contrairement aux études connexes qui étudient le CLS entre différents points de vue (Azarbonyad *et al.*, 2017; Rodriguez *et al.*, 2021), notre objectif n'est pas d'explorer l'utilisation de mots spécifiques, et nous n'évaluons pas notre méthode en fonction du classement des mots par la stabilité de leur sens. Nous voulons plutôt déterminer si les représentations vectorielles reflètent une plus grande similitude dans l'utilisation des mots au sein d'un point de vue qu'entre différentes positions (voir Figure 1). Nous explorons cette question en nous appuyant sur des ensembles de données annotés avec des informations sur la position. Auparavant, nous testons différents modèles contextuels dans une configuration où les données sont rares, afin de sélectionner un type de représentation robuste [2].

Notre objectif à long terme est de détecter les différences d'utilisation des mots entre les locuteurs d'une conversation, ce qui pourrait indiquer leur niveau d'alignement conceptuel (Stolk *et al.*, 2016) ; c'est-à-dire la mesure dans laquelle les participants au dialogue « veulent dire la même chose en utilisant les mêmes mots » (Schober, 2005). Avec cette étude, nous présentons un premier pas dans cette direction. Les représentations sensibles aux différences d'opinion pourraient être utiles pour identifier les désaccords et les désalignements dans le dialogue.

# 2 Méthodologie

## 2.1 Données

Nous utilisons des ensembles de données en anglais contenant des phrases étiquetées comme étant en faveur (FAVOR) ou contre (AGAINST) une cible spécifique. Nous excluons les phrases sans position claire (NONE), le cas échéant. **SemEval2016** (Mohammad *et al.*, 2016b,a) contient des tweets sur six cibles variées. Nous en extrayons 3 253 phrases [3]. **Covid19** (Glandt *et al.*, 2021) est un ensemble de données avec 3 918 tweets centrés sur quatre cibles liées à la pandémie de COVID-19. **P-Stance** (Li *et al.*, 2021) contient 21 574 tweets sur trois politiciens américains. Enfin, IBM-ArgQ-Rank-30kArgs (Gretz *et al.*, 2020), ci-après **ArgQ**, est une collection d'arguments sur 71 cibles. Nous

2. Notre code et nos données sont disponibles sur https://github.com/ainagari/1word2sides
3. Nous omettons la cible « Climate Change is a Real Concern » car elle ne contient que 26 tweets AGAINST.

utilisons 29 972 arguments qui ont une position claire (avec un score de confiance [4] supérieur à 0,6, d'après Bar-Haim *et al.* (2020)). Nous voulons organiser les données de manière à nous permettre de déterminer si des instances du même mot ont une similitude plus élevée au sein d'une position qu'entre les positions. À cette fin, nous prétraitons et organisons les données comme suit.

**Prétraitement :** L'ensemble de données ArgQ était à l'origine destiné à la détection de la qualité des arguments, et plusieurs arguments mentionnent explicitement leur position. Pour atténuer les biais potentiels que cela pourrait engendrer, nous appliquons une stratégie qui omet automatiquement cette partie d'une phrase. Si le début d'une phrase contient les mêmes mots que la cible (avec l'ajout facultatif de *not* et *n't*) et est suivi de *because* (*of*), *as*, *since*, d'une virgule ou d'un point, on omet la première partie de la phrase jusqu'à ce token (inclus) [5]. Cette procédure modifie 3 223 phrases. Ensuite, les phrases dans tous les ensembles de données sont tokenisées, annotées en parties du discours et lemmatisées avec la librairie `nltk`.

**Ensembles de phrases :** Pour une cible donnée, nous divisons aléatoirement les phrases de chaque position ($f$ (FAVOR) ou $a$ (AGAINST)) en deux ensembles de taille égale $P$ et $Q$. Avec ces ensembles, nous exécutons quatre comparaisons, deux intra-position : WITHIN-FAVOR ($P_f$ vs $Q_f$) et WITHIN-AGAINST ($P_a$ vs $Q_a$) ; et deux entre-positions : BETWEEN-1 ($P_f$ vs $Q_a$) et BETWEEN-2 ($P_a$ vs $Q_f$).

## 2.2 Représentations vectorielles

Nous voulons générer des représentations vectorielles pour des ensembles d'instances de mots tirés d'une position (par exemple, $P_f$). Par exemple, on veut obtenir une représentation du mot « woman » à partir de phrases en faveur du « mouvement féministe » (SemEval2016) et la comparer à la représentation de « woman » dans des phrases exprimant une position contre cette cible.

Dans la détection de CLS, les plongements statiques ont tendance à être plus performants que ceux contextualisés (Schlechtweg *et al.*, 2020). Une approche typique consiste à apprendre séparément les plongements statiques pour chaque période, corpus ou point de vue, puis les comparer soit en les alignant (Hamilton *et al.*, 2016) soit avec un approche basée sur les plus proches voisins (Gonen *et al.*, 2020). Dans ces études, même dans celles portant sur la détection de changement à court terme (Stewart *et al.*, 2017; Del Tredici *et al.*, 2019), il est courant de disposer d'un assez grande nombre d'instances d'un mot donné. Cependant, le nombre de phrases disponibles par mot dans un point de vue est limité dans nos données [6]. Nous expérimentons donc avec trois types différents de plongements contextualisés :

**Les plongements *À la carte* (ALC)** (Khodak *et al.*, 2018) ont été utilisées pour détecter des différences dans l'utilisation des mots selon les points de vue (Rodriguez *et al.*, 2021). Le modèle consiste à appliquer une transformation linéaire à la moyenne des plongements pré-entraînés des mots du contexte du mot cible. Nous utilisons un modèle ALC reposant sur des plongements GloVe 300*d* (Pennington *et al.*, 2014) formés sur 840 $\times 10^9$ tokens de Common Crawl.

4. Ce score reflète la mesure dans laquelle les annotateurs sont d'accord sur la position d'un argument. Il est calculé comme une moyenne pondérée des décisions des annotateurs et il varie de 0 à 1.

5. Par exemple, on retient seulement la partie en italique pour la phrase « Homeschooling should not be banned because *it is a right for parents to educate their children in their comfort of home* . » (pour la cible « Homeschooling should be banned »).

6. Par exemple, Schlechtweg *et al.* (2020) a une moyenne de 788 instances par lemme et période de temps. Dans nos données, le nombre moyen d'instances d'un mot dans un côté d'une comparaison est de 14.

**Context2vec (c2v)** (Melamud *et al.*, 2016) qui apprend simultanément la représentation d'un mot cible et, à l'aide d'un biLSTM, du contexte qui l'entoure : il est optimisé pour que les représentations du mot et de son contexte, plongées dans le même espace, soient similaires. Nous utilisons un modèle 600$d$ entraîné sur le corpus ukWaC (Baroni *et al.*, 2009).

**BERT :** (Devlin *et al.*, 2019) Nous utilisons des représentations contextualisées générées avec le modèle 768$d$ `bert-base-uncased`.

Nous notons $V_P$ le vocabulaire d'un ensemble de phrases $P$. Nous incluons dans le vocabulaire tous les noms et les verbes apparaissant dans au moins trois phrases différentes de $P$. Dans les tweets, les mentions et les hashtags sont traités comme des noms. Les mots vides sont exclus. Nous traitons toutes les instances d'un lemme avec une catégorie grammaticale spécifique comme le même mot. Nous extrayons une représentation $\mathbf{w}_P$ pour chaque mot $w$ dans $V_P$. Pour c2v et BERT, cela se fait en faisant la moyenne des représentations de toutes les instances $w$ de $P$.

## 2.3 Tester les représentations

Nous identifions d'abord les représentations les mieux adaptées pour refléter la similarité sémantique lexicale entre de petits ensembles de phrases. En suivant Schlechtweg & Schulte im Walde (2020), nous utilisons SemCor (Miller *et al.*, 1993), un corpus annoté en sens, pour créer un ensemble de données qui simule le changement sémantique. Nous contrôlons en outre le nombre de phrases disponible pour chaque lemme. Le jeu de données est composé de 576 lemmes : 245 noms, 241 verbes, 69 adjectifs et 21 adverbes. Pour chaque lemme, nous avons deux ensembles de 25 instances chacun, $P$ et $Q$. Pour simuler des situations de données, nous créons des sous-ensembles de taille $X$ de $P$ et $Q$ ($P_X$ , $Q_X$). Nous expérimentons différentes valeurs de $X$. Comme dans Schlechtweg & Schulte im Walde (2020), nous déterminons la « vraie » distance sémantique entre deux groupes $P_X$ et $Q_X$ en calculant la divergence de Jensen-Shannon (JSD) entre leurs distributions de sens.

Les prédictions de similarité pour un mot $w$ sont obtenues en calculant simplement la similarité cosinus entre les représentations de ce lemme dans chaque ensemble de phrases, $cos(\mathbf{w}_{P_X}, \mathbf{w}_{Q_X})$. Nous rapportons le coefficient de corrélation $\tau$-b de Kendall entre JSD et les similarités prédites par chaque type de représentation. Les résultats de cette expérience sont présentés dans la section 3.1.

## 2.4 Calcul de similarité et évaluation

**Calcul de similarité :** Pour calculer la similarité globale de l'utilisation des mots pour une comparaison entre deux ensembles de phrases $P$ et $Q$, nous identifions d'abord les mots communs aux deux ensembles, $V_P \cap V_Q$. Cependant, $V_P \cap V_Q$ contient des mots qui ne sont pas nécessairement liés à la cible en question. On calcule donc une similarité basée uniquement sur un sous-ensemble de $V_P \cap V_Q$, appelé $V_{PQ}$. Le score de similarité est la similarité cosinus moyenne de tous les mots dans $V_{PQ}$ :

$$sim(P, Q) = \frac{\sum_{w \in V_{PQ}} cos(\mathbf{w}_P, \mathbf{w}_Q)}{|V_{PQ}|} \quad (1)$$

Cette mesure de similarité vise à refléter dans quelle mesure les mots sont utilisés de la même manière et dans le même sens dans deux ensembles de phrases. Nous expérimentons avec trois définitions de

$V_{PQ}$. Dans chacune, nous veillons à utiliser le même nombre de mots pour les quatre comparaisons qui sont faites au sein d'une cible. Dans *all*, on inclut les $k$ premiers mots les plus fréquents dans $V_P \cap V_Q$, où $k$ correspond à la plus petite taille de $V_P \cap V_Q$ disponible pour cette cible. La fréquence est déterminée à partir de l'union des phrases dans $P$ et $Q$. Nous utilisons également les 10 premiers mots de $V_P \cap V_Q$ avec les scores tf-idf les plus élevés dans cette cible (*tf-idf*). Les scores tf-idf sont calculés sur l'ensemble des jeux de données, en traitant toutes les phrases concernant la même cible comme un seul document. Enfin, on utilise aussi les 10 mots de $V_P \cap V_Q$ avec le plus petit poids de tf-idf (*rev-tf-idf*). Ce sous-ensemble contient des mots qui sont moins pertinents pour la cible, et donc nous nous attendons à ce que les similarités BETWEEN- et WITHIN- aient des valeurs plus proches. Notons que 25% des comparaisons (dans SemEval2016 et ArgQ) ont moins de 20 mots en commun. Dans ces cas, les valeurs *tf-idf* et *rev-tf-idf* sont partiellement calculées avec les mêmes mots.

**Évaluation :** Nous nous attendons à ce que les comparaisons de type WITHIN présentent une similarité moyenne plus élevée que les comparaisons BETWEEN. Pour mesurer à quel point cela est vrai, nous utilisons la précision par paires : nous vérifions pour combien de paires (WITHIN, BETWEEN) la comparaison de type BETWEEN a une similarité plus faible. Avec 4 comparaisons par cible, nos expériences portent sur un total de 332 paires (WITHIN, BETWEEN). Les résultats sur les données de position sont présentées dans la section 3.2.

# 3 Résultats

## 3.1 Sélection d'un type de représentation

Les résultats sur SemCor sont présentés sur la Figure 2. Dans les graphiques *a* et *b*, nous voyons les corrélations obtenues par les différents types de représentations sur différentes quantités de données ($X$). Naturellement, les performances sont moins bonnes avec des valeurs faibles de $X$. C'est notamment le cas de ALC, qui à $X = 25$ continue à s'améliorer. Dans le cas de c2v et BERT, cependant, nous n'observons pas de d'amélioration significative après $X = 10$. Dans ce contexte de données rares, les performances des plongements ALC sont bien inférieures à celles de c2v et BERT. Dans l'ensemble, les représentations BERT de la 10ème couche fonctionnent le mieux. On utilise donc des plongements de cette couche pour nos expériences sur les données de points de vue. Nous examinons également les performances des deux meilleurs modèles (c2v et la 10ème couche dans BERT) par PoS (figures *c* et *d*) : nous constatons que les noms et les verbes, qui sont les catégories grammaticales incluses dans nos expériences de position, sont généralement mieux représentés.

## 3.2 Résultats sur la position

Les précisions par paires obtenues avec la 10ème couche BERT avec différentes définitions de $V_{PQ}$ se trouvent dans la Table 1 (gauche). Nous voyons que, en particulier pour *all* et *tf-idf*, la précision par paires est remarquablement élevée dans tous les ensembles de données. Cela montre que les représentations de mots contextualisées de BERT reflètent des différences dans la façon dont les mots sont utilisés entre deux postures opposées.

Lors de l'utilisation des 10 mots avec le tf-idf le plus faible (*rev-tf-idf*), les performances diminuent,

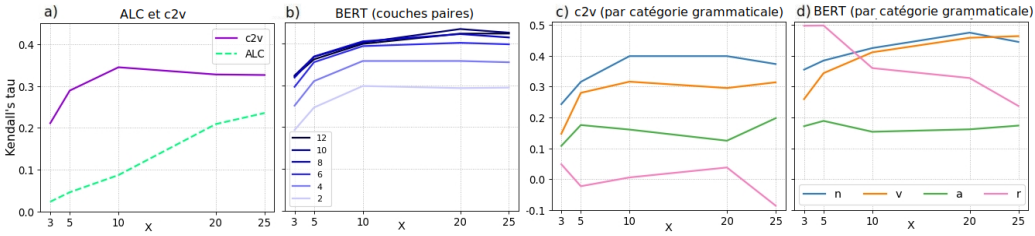


FIGURE 2 – *a* and *b* : $\tau$ de Kendall obtenu par différentes représentations vectorielles sur SemCor. Nous n'incluons que des couches paires pour BERT pour une meilleure lisibilité. *c* et *d* : Performances de c2v et BERT (10ème couche) par catégorie grammaticale.

| **Dataset** | ***all*** | **tf-idf** | **rev-tf-idf** |
|---|---|---|---|
| SemEval2016 | 0.90 | 0.85 | 0.60 |
| Covid19 | 0.88 | 0.81 | 0.50 |
| P-stance | 1.00 | 1.00 | 0.83 |
| ArgQ | 1.00 | 0.98 | 0.95 |
| Global | 0.99 | 0.96 | 0.90 |

| | ***all*** | **tf-idf** | **rev-tf-idf** |
|---|---|---|---|
| a) W vs W | 0.013 | 0.010 | 0.023 |
| b) B vs B | 0.013 | 0.010 | 0.023 |
| c) W vs B | 0.047 | 0.027 | 0.041 |

TABLE 1 – À gauche : Précision par paires par ensemble de données avec différents $V_{PQ}$. *Global* correspond à tous les jeux de données réunis. À droite : Différences de similarité entre comparaisons.

mais elles restent élevées pour les datasets P-Stance et ArgQ. Nous exécutons des tests du $\chi^2$ pour la qualité d'ajustement sur les prédictions *rev-tf-idf* pour déterminer leur probabilité sous l'hypothèse nulle ($H_0$ : préc. $= 0.5$). Les valeurs-p sont significatives pour tous les jeux de données ensemble ($p < 0.001$) mais pas pour l'ensemble de toutes les données Twitter ($p = 0.08$, $\alpha = 0.05$). Il semble que les représentations BERT encodent, dans une certaine mesure, les différences dans les mots qui sont moins pertinents pour la cible. Cependant, si pour une raison quelconque, tous les mots ne peuvent pas être utilisés, (s'il y en a trop), alors il est préférable de sélectionner soigneusement un sous-ensemble (par exemple avec tf-idf).

Nous examinons également les mots qui ont les similitudes les plus élevées et les plus faibles dans les comparaisons BETWEEN. Les mots qui sont utilisés le plus différemment entre les positions tendent à être des noms qui sont au centre du sujet (par exemple « religion » dans « athéisme »), tandis que les mots les plus similaires sont souvent non-thématiques (« man » ou « take »).

Nous étudions l'ampleur des différences de similarité entre les comparaisons WITHIN (W) et BETWEEN (B) en examinant les différences de similarité (en valeur absolue) entre les paires de comparaison : a) entre WITHIN-FAVOR et WITHIN-AGAINST (W vs W), b) entre BETWEEN-1 et BETWEEN-2 (B vs B), et c) la différence moyenne trouvée dans les quatre appariements WITHIN vs BETWEEN (W vs B). Nous nous attendons à ce que ce dernier ait une plus grande différence de similarité que a) et b), où les comparaisons sont du même type. Les résultats sont présentés dans la Table 1 (droite). Nous rapportons la moyenne de ces valeurs sur toutes les données. Les différences de similarité sont assez faibles dans l'ensemble, ce qui indique que le contraste (c'est-à-dire la mesure dans laquelle les comparaisons WITHIN affichent une plus grande similarité que les comparaisons BETWEEN) est subtile. Les valeurs sont cependant entre 1,8 et 3,6 fois plus grandes pour les paires de comparaison W vs B. Pour toutes les définitions $V_{PQ}$, les valeurs de différence des paires de comparaison sont

significativement éloignées de celles en a) et b) ($p < 0{,}001$). [7]

# 4 Conclusion et travaux futurs

Nous avons montré que les représentations de mots BERT sont sensibles à l'opinion exprimée dans les phrases dont ils sont dérivés. Les différences de similitude trouvées entre les positions concordantes et contradictoires sont petites, mais significatives ; et les mots avec les différences les plus élevées ont tendance à être au cœur du sujet. Notre approche peut servir à identifier les points de divergence par rapport à une cible, et elle peut être utile pour la détection de position et l'analyse des débats. Nos expériences sur SemCor fournissent des informations précieuses sur la quantité suffisante d'instances de mots nécessaire à l'obtention de représentations de qualité, ce qui est pertinent pour étudier le CLS à faibles ressources et, plus généralement, pour dériver des vecteurs de mots à partir de peu de données.

Dans nos futurs travaux, nous prévoyons d'appliquer cette méthodologie au dialogue. Les ensembles $P$ et $Q$ correspondraient aux énoncés des participants à une conversation. La mesure de similarité agirait comme une approximation de l'alignement conceptuel ou de position entre les deux participants, indiquant si les locuteurs partagent des opinions et utilisent des mots de manière similaire.

# Remerciements

Nous remercions les relecteurs anonymes pour leurs commentaires utiles. Ce travail a été soutenu par la chaire de recherche Télécom Paris sur la Science des Données et Intelligence Artificielle pour l'Industrie et les Services Numériques (DSAIDIS).

# Références

AZARBONYAD H., DEHGHANI M., BEELEN K., ARKUT A., MARX M. & KAMPS J. (2017). Words Are Malleable : Computing Semantic Shifts in Political and Media Discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, p. 1509–1518, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3132847.3132878.

BAR-HAIM R., EDEN L., FRIEDMAN R., KANTOR Y., LAHAV D. & SLONIM N. (2020). From Arguments to Key Points : Towards Automatic Argument Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4029–4039, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.371.

BARONI 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**(3), 209–226.

7. Selon les tests de Wilcoxon ou de Student appariés, dépendant de la normalité des données (déterminée par les tests de Shapiro-Wilk).

DEL TREDICI M., FERNÁNDEZ R. & BOLEDA G. (2019). Short-Term Meaning Shift : A Distributional Exploration. In *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)*, p. 2069–2075, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1210.

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, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GARÍ SOLER A., LABEAU M. & CLAVEL C. (2022). One word, two sides : Traces of stance in contextualized word representations. In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3950–3959, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

GARIMELLA A., MIHALCEA R. & PENNEBAKER J. (2016). Identifying Cross-Cultural Differences in Word Usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 674–683, Osaka, Japan : The COLING 2016 Organizing Committee.

GLANDT K., KHANAL S., LI Y., CARAGEA D. & CARAGEA C. (2021). Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1596–1611, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.127.

GONEN H., JAWAHAR G., SEDDAH D. & GOLDBERG Y. (2020). Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 538–555, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.51.

GRETZ S., FRIEDMAN R., COHEN E., TOLEDO A., LAHAV D., AHARONOV R. & SLONIM N. (2020). A Large-Scale Dataset for Argument Quality Ranking : Construction and Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 7805–7813. DOI : 10.1609/aaai.v34i05.6285.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1489–1501, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1141.

KHODAK M., SAUNSHI N., LIANG Y., MA T., STEWART B. & ARORA S. (2018). A La Carte Embedding : Cheap but Effective Induction of Semantic Feature Vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 12–22, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1002.

LI Y., SOSEA T., SAWANT A., NAIR A. J., INKPEN D. & CARAGEA C. (2021). P-Stance : A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 2355–2365, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.208.

MELAMUD O., GOLDBERGER J. & DAGAN I. (2016). context2vec : Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Compu-*

*tational Natural Language Learning*, p. 51–61, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-1006.

MILLER G. A., LEACOCK C., TENGI R. & BUNKER R. T. (1993). A Semantic Concordance. In *Human Language Technology : Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

MOHAMMAD S., KIRITCHENKO S., SOBHANI P., ZHU X. & CHERRY C. (2016a). A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 3945–3952, Portorož, Slovenia : European Language Resources Association (ELRA).

MOHAMMAD S., KIRITCHENKO S., SOBHANI P., ZHU X. & CHERRY C. (2016b). SemEval-2016 Task 6 : Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 31–41, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1003.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1162.

RODRIGUEZ P. L., SPIRLING A. & STEWART B. M. (2021). *Embedding Regression : Models for Context-Specific Description and Inference*. Rapport interne, Working Paper Vanderbilt University.

SCHLECHTWEG D., HÄTTY A., DEL TREDICI M. & SCHULTE IM WALDE S. (2019). A Wind of Change : Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 732–746, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1072.

SCHLECHTWEG D., MCGILLIVRAY B., HENGCHEN S., DUBOSSARSKY H. & TAHMASEBI N. (2020). SemEval-2020 Task 1 : Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, p. 1–23, Barcelona (online) : International Committee for Computational Linguistics. DOI : 10.18653/v1/2020.semeval-1.1.

SCHLECHTWEG D. & SCHULTE IM WALDE S. (2020). Simulating Lexical Semantic Change from Sense-Annotated Data. In A. RAVIGNANI, C. BARBIERI, M. MARTINS, M. FLAHERTY, Y. JADOUL, E. LATTENKAMP, H. LITTLE, K. MUDD & T. VERHOEF, Éds., *The Evolution of Language : Proceedings of the 13th International Conference (EvoLang13)*. DOI : 10.17617/2.3190925.

SCHOBER M. F. (2005). Conceptual alignment in conversation. *Other minds : How humans bridge the divide between self and others*, p. 239–252.

STEWART I., ARENDT D., BELL E. & VOLKOVA S. (2017). Measuring, Predicting and Visualizing Short-Term Change in Word Representation and Usage in VKontakte Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, **11**(1), 672–675.

STOLK A., VERHAGEN L. & TONI I. (2016). Conceptual alignment : How brains achieve mutual understanding. *Trends in cognitive sciences*, **20**(3), 180–191.

TAHMASEBI N., BORIN L. & JATOWT A. (2021). Survey of computational approaches to lexical semantic change detection. In N. TAHMASEBI, L. BORIN, A. JATOWT, Y. XU & S. HENGCHEN, Éds., *Computational approaches to semantic change*. Language Science Press. DOI : 10.5281/zenodo.5040302.

YIN Z., SACHIDANANDA V. & PRABHAKAR B. (2018). The global anchor method for quantifying linguistic shifts and domain adaptation. *Advances in neural information processing systems*, **31**.