

What if the whole is greater than the sum of the parts? Modelling Complex (Multiword) Expressions

Aline Villavicencio

Department of Computer Science, University of Sheffield
Regent Court (DCS)
211 Portobello
Sheffield, S1 4DP, UK

Abstract

Multiword Expressions (MWEs) such as idioms (make ends meet), light verb constructions (give a sigh), verb particle constructions (shake up) and noun compounds (loan shark), are an integral part of the mental lexicon of native speakers often used to express complex ideas in a simple and conventionalised way accepted by a given linguistic community. As they may display a wealth of idiosyncrasies, from lexical, syntactic and semantic to statistical, they have represented a real challenge for natural language processing. However, their accurate integration has the potential for improving the precision, naturalness and fluency of downstream tasks like text simplification. In this paper I discuss some advances in the identification and modelling of MWEs, concentrating on techniques for identifying their degree of idiomaticity and approximating their meaning. One of the challenges is that their interpretation often needs more knowledge than can be gathered from their individual components and their combinations to differentiate combinations whose meaning can be (partly) inferred from their parts (as apple juice: juice made of apples) from those that cannot (as dark horse: an unknown candidate who unexpectedly succeeds). In particular, the paper presents results obtained with the use of contextualised word representation models, which have been successfully used for capturing different word usages, and therefore could provide an attractive alternative for representing idiomaticity in language.

Keywords

Multiword Expressions, Idiomaticity, Lexical Simplification, Contextualised Word Embeddings

1. Introduction

Human language is a powerful means for communicating ideas, and concepts, requests and desires, for transmitting folklore, tales, history and scientific knowledge, on both an individual and a global scale. It is expressive allowing complex ideas to be transmitted both formally and informally, in scientific and in colloquial settings; it is creative and dynamic incorporating new concepts, words and usages; flexible and ambiguous allowing for irony, jokes, idioms and metaphors (e.g. *a blanket of snow*, *French kiss*, *rocket science*). However, such a large, rich and complex system may also challenge humans and create barriers for a clear understanding of the message to be communicated. For instance, a lack of vocabulary or of a particular word usage

Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021), co-located with SEPLN 2021, September 21st, 2021 (Online). Saggion, H., Štajner, S. and Ferrés, D. (Eds).


✉ a.villavicencio@sheffield.ac.uk (A. Villavicencio)

🌐 <https://sites.google.com/view/alinev> (A. Villavicencio)

🆔 0000-0002-3731-9168 (A. Villavicencio)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

may hinder the full understanding of a message. This is a common situation for children learning a language [1], for adult native speakers learning a new domain, and often faced by second language learners [2]. Health factors have also been linked with an impact in language abilities, including in clinical conditions like aphasia [3] and dyslexia [4] and autism [5]. Socio-economic conditions and lack of access to schooling may also result in low language abilities and even illiteracy [6], which, in many contexts, may affect a large part of the population, as, for instance, in Brazil, where according to social indicators, in 2012 6% of the Brazilian population aged 15 or over were fully illiterate, among 27% who were functionally illiterate.¹ Therefore, making texts accessible for different target groups can be viewed as a crucial task not only for leading to a better understanding, but also for ensuring a better quality of life.

In fact, there have been considerable research efforts into how to simplify texts [7, 8, 9, 10], analysing their readability levels [11, 12, 13], creating resources [14, 15, 16], proposing new techniques [17, 14] and analysing the suitability of evaluation frameworks [11, 15, 18]. These have been applied for languages that in addition to English, include Spanish [19, 20], Portuguese [6, 12, 21, 22], Italian [23] and Basque [24].

One approach to text simplification targets lexical substitutions [7], where words identified as difficult are replaced by easier synonyms, according to a given simplicity measure, such as word frequency, polysemy and length. However, in addition to considering words individually, lexical substitution needs to take into account the occurrence of Multiword Expressions (MWEs) in sentences, including compound nouns (like *loan shark*²) and idioms (like *let the cat out of the bag*³) since their meanings may not be related to the meanings of their individual components, and replacing one of their components in isolation may result in a nonsensical simplification or even in loss of the original meaning (e.g. *loan fish* instead of *loan shark*). In particular, MWEs are also challenging since they display many idiosyncrasies, from lexical to semantic and statistical [25], and these may have an effect on the readability of a sentence, and are particularly problematic for non-native speakers of a language since these idiosyncrasies are often unpredictable, and may lead to markedness (e.g. *strong coffee* vs. *?powerful coffee*).

Indeed, the positive impact of explicitly handling MWEs has been discussed for a number of tasks and applications like machine translation [26], information retrieval [27] and parsing [28], as discussed by Constant et al [29]. Although text simplification has received considerable attention, studies focusing on MWEs are still scarce and there is a need for further investigation into potential processing overheads involved in MWEs (if any) for different groups, and how they can best be handled for text simplification. In what follows, I discuss some of the challenges of MWEs and some techniques for detecting their idiomaticity.

2. Multiword Expressions and Idiomaticity

MWEs have been defined as word sequences, not necessarily adjacent, that are recurrent, act as a single unit at some level of linguistic analysis [30], and whose interpretation crosses word boundaries [31]. MWEs include, besides compound nouns (*cheese knife, software engineering*)

¹<http://www.ipm.org.br/>

²Meaning a person who lends money at very high interest rates.

³Meaning to reveal a secret.

and idioms (*make ends meet, kick the bucket*), verb-particle constructions (*break down, carry on*), determinerless PPs (*in hospital*), collocations (*strong coffee, heave baggage*), support verbs (*give sigh, take shower*), among others. The importance of MWEs for a simplification system can be gauged from estimates about their frequency in language: for Biber et al. [32] they correspond to between 30% and 45% of spoken English and 21% of academic prose, while for Jackendoff [33] they have the same order of magnitude as the number of single words in a speaker’s mental lexicon. They have also been found to have faster processing times compared to non-MWEs (compositional novel sequences) [34, 35].

Their computational treatment involves steps like discovering their occurrence in sentences, determining how idiomatic they are (as a type in general and as an MWE token in a particular sentence), and approximating their meaning. For MWE discovery, methods for determining if a given sequence of words forms an MWE or not have often been based on the statistical markedness of their recurrence, using association and entropic measures calculated from corpus counts [36], and on morphosyntactic patterns commonly associated to MWEs [37, 38]. These have been applied to a variety of MWE types and languages and have been extensively discussed [39, 36, 40].

For determining the meaning of a combination of words, one common strategy is to derive it from the meanings of the parts. However, for idiomatic cases this approach will lead to an unrelated meaning being derived. This difference between the meaning of the MWE and the meanings of the components can be used as the basis for determining how idiomatic a given MWE is [41, 42, 43, 44]. The assumption is that the closer the meaning of the MWE is to the meaning of the components, the more compositional it is, and the more they differ, the more idiomatic the MWE is. This approach has been used with word embeddings, using the proximity in a multidimensional space between the embedding of the MWE and the embedding of the components combined by operations like vector addition as semantic proximity. The results obtained with static word embeddings like word2vec [45] and GloVe [46] for MWE type idiomaticity detection for instance for Noun Compounds in languages like English, French and Portuguese have been strongly correlated with human judgments about idiomaticity [44].

For token idiomaticity detection, the challenge is to decide if a potentially ambiguous MWE is used literally or idiomatically in a given sentence (e.g. *big fish* literally as a large aquatic animal or idiomatically as an important person). For this task, contextualised word embeddings like BERT [47] and ELMo [48] may be an attractive alternative, as they seem to represent words more accurately than static word embeddings like GloVe, being able to encode different usages of a given word that appear to form clusters that seem related to its various senses [49]. However, analyses of whether and to what extent idiomaticity in MWEs is accurately incorporated by word representation models have recently reported mixed results. On the one hand, in an analysis of different classifiers initialised with contextualised and non-contextualised embeddings evaluated in five tasks related to lexical composition (including the literality of NCs) contextualised models, especially BERT, obtained the best performance across all tasks [50]. On other analyses, static models like *word2vec* had better performance than contextualised models [51, 52]. These mixed results suggest that a controlled evaluation setup may be needed to obtain comparable results across models and languages.

In recent work a set of probing tasks were defined to assess the representation of noun compounds (NCs) in vector space models in two languages, English and French. The probes

took into account the NCs and their paraphrases in contexts that involved minimal modifications and were used to compare the idiomatic and literal representations of a given NC [53]. The goal was to assess if word embeddings with different levels of contextualisation would be sensitive to changes in idiomaticity caused by different paraphrases in naturalistic sentences and in context neutral sentences.

A second set of probes was defined to assess if the models were more sensitive to idiomaticity at type or at token level [54], and whether contextualised models could reach the performance obtained by static embeddings for type idiomaticity detection [44]. The results obtained in these evaluations suggest that these models are still not sensitive enough to detect idiomaticity accurately. Moreover, even if contextualised embeddings outperform static models in many tasks, they still have lower performance for idiomaticity detection. Further analyses that take into account more fine-grained sense distinctions for MWEs, allowing for polysemy in literal and in idiomatic senses, have revealed that fine-tuning improves performance, both for MWE discovery in sentences and for sense identification [55]. This analysis is part of SEMEVAL 2022 Task⁴, and includes datasets for English, Portuguese and Galician.

For downstream tasks like text simplification, these results mean that systems that adopt state-of-the-art pre-trained models as they are, will lack accuracy when faced with MWEs, interpreting an idiomatic MWE like *eager beaver* in a sentence such as *When she first started working she was a real **eager beaver***. as more similar to *anxious castor* than to its actual synonym of *hardworking person*.

3. Conclusions

In this paper I discussed MWEs and some of the challenges they pose for tasks like text simplification. In particular I concentrated on recent work on idiomaticity detection using state-of-the-art language models. The results obtained suggest that there are still advances to be made for more accurate representation of MWE idiomaticity. However, given the prevalence of MWE in natural languages, and their role in both general and technical domains, approaches for text simplification, especially lexical simplification, need to take them into account for more precise and understandable results. Moreover, additional research on the impact of MWEs for different groups of speakers can shed light into any additional complexity, given their advantageous processing for native speakers, and their often opaque meaning for non-native speakers.

Acknowledgments

I want to thank the many co-authors who have collaborated in this work, including Marco Idiart, Carlos Ramisch, Carolina Scarton, Harish Tayyar Madabushi, Tiago Vieira, Marcos Garcia, Rodrigo Wilkens, Leonardo Zilio, Renata Ramisch and Silvio Cordeiro (in no particular order). This research has been partly funded by EPSRC project MIA.

⁴<https://sites.google.com/view/semEval2022task2-idiomaticity>

References

- [1] J. De Belder, M.-F. Moens, Text simplification for children, ACM; New York, 2010, pp. 19–26.
- [2] G. H. Paetzold, Lexical simplification for non-native english speakers, 2016.
- [3] J. A. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, J. Tait, Simplifying text for language-impaired readers, in: EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway, The Association for Computer Linguistics, 1999, pp. 269–270. URL: <https://aclanthology.org/E99-1042/>.
- [4] L. Rello, R. Baeza-Yates, The effect of font type on screen readability by people with dyslexia, ACM Trans. Access. Comput. 8 (2016) 15:1–15:33. URL: <https://doi.org/10.1145/2897736>. doi:10.1145/2897736.
- [5] R. Evans, C. Orasan, I. Dornescu, An evaluation of syntactic simplification rules for people with autism, in: S. Williams, A. Siddharthan, A. Nenkova (Eds.), Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR@EACL 2014, Gothenburg, Sweden, April 27, 2014, Association for Computational Linguistics, 2014, pp. 131–140. URL: <https://doi.org/10.3115/v1/W14-1215>. doi:10.3115/v1/W14-1215.
- [6] S. Aluísio, C. Gasperin, Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts, in: Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 46–53. URL: <https://aclanthology.org/W10-1607>.
- [7] A. Siddharthan, An architecture for a text simplification system, in: 2002 Language Engineering Conference (LEC 2002), 13-15 December 2002, Hyderabad, India, IEEE Computer Society, 2002, p. 64. URL: <https://doi.org/10.1109/LEC.2002.1182292>. doi:10.1109/LEC.2002.1182292.
- [8] M. Shardlow, A comparison of techniques to automatically identify complex words, in: 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Student Research Workshop, 4-9 August 2013, Sofia, Bulgaria, The Association for Computer Linguistics, 2013, pp. 103–109. URL: <https://aclanthology.org/P13-3015/>.
- [9] S. Štajner, H. Saggion, Data-driven text simplification, in: Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 19–23. URL: <https://aclanthology.org/C18-3005>.
- [10] F. Alva-Manchego, C. Scarton, L. Specia, Data-driven sentence simplification: Survey and benchmark, Computational Linguistics 46 (2020) 135–187. URL: <https://aclanthology.org/2020.cl-1.4>. doi:10.1162/coli_a_00370.
- [11] S. Štajner, R. Mitkov, H. Saggion, One step closer to automatic evaluation of text simplification systems, in: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 1–10. URL: <https://aclanthology.org/W14-1201>. doi:10.3115/v1/W14-1201.

- [12] J. A. Wagner Filho, R. Wilkens, A. Villavicencio, Automatic construction of large readability corpora, in: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 164–173. URL: <https://aclanthology.org/W16-4119>.
- [13] M. Maddela, W. Xu, A word-complexity lexicon and a neural readability ranking model for lexical simplification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3749–3760. URL: <https://aclanthology.org/D18-1410>. doi:10.18653/v1/D18-1410.
- [14] W. Coster, D. Kauchak, Simple english wikipedia: A new text simplification task, in: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers, The Association for Computer Linguistics, 2011, pp. 665–669. URL: <https://aclanthology.org/P11-2117/>.
- [15] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, L. Specia, ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4668–4679. URL: <https://aclanthology.org/2020.acl-main.424>. doi:10.18653/v1/2020.acl-main.424.
- [16] G. Paetzold, L. Specia, SemEval 2016 task 11: Complex word identification, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 560–569. URL: <https://aclanthology.org/S16-1085>. doi:10.18653/v1/S16-1085.
- [17] K. Woodsend, M. Lapata, Wikisimple: Automatic simplification of wikipedia articles, in: W. Burgard, D. Roth (Eds.), Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011, AAAI Press, 2011. URL: <http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3505>.
- [18] F. Alva-Manchego, C. Scarton, L. Specia, The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification, Computational Linguistics (2021). arXiv:https://direct.mit.edu/coli/article/doi/10.1162/coli_a00418/106930/The-Un-Suitability-of-Automatic-Evaluation-Metrics.
- [19] S. Štajner, I. Calixto, H. Saggion, Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 618–626. URL: <https://aclanthology.org/R15-1080>.
- [20] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making it simplext: Implementation and evaluation of a text simplification system for spanish, ACM Trans. Access. Comput. 6 (2015). URL: <https://doi.org/10.1145/2738046>. doi:10.1145/2738046.
- [21] N. Hartmann, G. H. Paetzold, S. Aluísio, SIMPLEX-PB 2.0: A reliable dataset for lexical simplification in Brazilian Portuguese, in: Proceedings of the The Fourth Widening Natural Language Processing Workshop, Association for Computational Linguistics, Seattle, USA, 2020, pp. 18–22. URL: <https://aclanthology.org/2020.winlp-1.6>. doi:10.18653/v1/2020.winlp-1.6.
- [22] S. Evaldo Leal, J. M. Munguba Vieira, E. dos Santos Rodrigues, E. Nogueira Teixeira, S. Aluí-

- sio, Using eye-tracking data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5821–5831. URL: <https://aclanthology.org/2020.coling-main.512>. doi:10.18653/v1/2020.coling-main.512.
- [23] G. Barlacchi, S. Tonelli, ERNESTA: A sentence simplification tool for children’s stories in italian, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II, volume 7817 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 476–487. URL: https://doi.org/10.1007/978-3-642-37256-8_39. doi:10.1007/978-3-642-37256-8_39.
- [24] I. Gonzalez-Dios, M. J. Aranzabe, A. D. de Ilarraza, The corpus of basque simplified texts (CBST), *Lang. Resour. Evaluation* 52 (2018) 217–247. URL: <https://doi.org/10.1007/s10579-017-9407-6>. doi:10.1007/s10579-017-9407-6.
- [25] T. Baldwin, S. N. Kim, Multiword expressions, in: N. Indurkha, F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, Second Edition, Chapman and Hall/CRC, 2010, pp. 267–292. URL: <http://www.crcnetbase.com/doi/abs/10.1201/9781420085938-c12>.
- [26] M. Carpuat, M. T. Diab, Task-based evaluation of multiword expressions: a pilot study in statistical machine translation, in: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, Proceedings, June 2-4, 2010, Los Angeles, California, USA, The Association for Computational Linguistics, 2010, pp. 242–245. URL: <https://aclanthology.org/N10-1029/>.
- [27] O. Acosta, A. Villavicencio, V. Moreira, Identification and treatment of multiword expressions applied to information retrieval, in: *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 101–109. URL: <https://aclanthology.org/W11-0815>.
- [28] M. Constant, J. Nivre, A transition-based system for joint lexical and syntactic analysis, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 161–171. URL: <https://aclanthology.org/P16-1016>. doi:10.18653/v1/P16-1016.
- [29] M. Constant, G. Eryiğit, J. Monti, L. van der Plas, C. Ramisch, M. Rosner, A. Todirascu, Survey: Multiword expression processing: A Survey, *Computational Linguistics* 43 (2017) 837–892. URL: <https://aclanthology.org/J17-4005>. doi:10.1162/COLI_a_00302.
- [30] N. Calzolari, C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, A. Zampolli, Towards best practice for multiword expressions in computational lexicons., in: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain, 2002. URL: <http://www.lrec-conf.org/proceedings/lrec2002/pdf/259.pdf>, aCL Anthology Identifier: L02-1259.
- [31] I. A. Sag, T. Baldwin, F. Bond, A. A. Copestake, D. Flickinger, Multiword expressions: A pain in the neck for NLP, in: A. F. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002, Proceedings, volume 2276 of *Lecture Notes in Computer Science*, Springer, 2002,

- pp. 1–15. URL: https://doi.org/10.1007/3-540-45715-1_1. doi:10.1007/3-540-45715-1_1.
- [32] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, Longman Grammar of Spoken and Written English, Pearson Education, 1999.
- [33] R. Jackendoff, *The Architecture of the Language Faculty*, volume 28, MIT Press, 1997.
- [34] M. H. Christiansen, I. Arnon, E. Lieven, A. Wray, Multiword sequences as building blocks for language: Insights into first and second language learning, in: M. Knauff, M. Pauen, N. Sebanz, I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society, CogSci 2013*, Berlin, Germany, July 31 - August 3, 2013, cognitivesciencesociety.org, 2013. URL: <https://mindmodeling.org/cogsci2013/papers/0029/index.html>.
- [35] A. Siyanova-Chanturia, K. Conklin, N. Schmitt, Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers, *Second Language Research* 27 (2011) 251–272. URL: <https://doi.org/10.1177/0267658310382068>. doi:10.1177/0267658310382068. arXiv:<https://doi.org/10.1177/0267658310382068>.
- [36] P. Pecina, Lexical association measures and collocation extraction, *Lang. Resour. Evaluation* 44 (2010) 137–158. URL: <https://doi.org/10.1007/s10579-009-9101-4>. doi:10.1007/s10579-009-9101-4.
- [37] J. S. Justeson, S. M. Katz, Technical terminology: some linguistic properties and an algorithm for identification in text, *Nat. Lang. Eng.* 1 (1995) 9–27. URL: <https://doi.org/10.1017/S1351324900000048>. doi:10.1017/S1351324900000048.
- [38] T. Baldwin, Deep lexical acquisition of verb-particle constructions, *Comput. Speech Lang.* 19 (2005) 398–414. URL: <https://doi.org/10.1016/j.csl.2005.02.004>. doi:10.1016/j.csl.2005.02.004.
- [39] C. Ramisch, *Multiword Expressions Acquisition - A Generic and Open Framework*, Theory and Applications of Natural Language Processing, Springer, 2015. URL: <https://doi.org/10.1007/978-3-319-09207-2>. doi:10.1007/978-3-319-09207-2.
- [40] N. Schneider, D. Hovy, A. Johannsen, M. Carpuat, SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM), in: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, Association for Computational Linguistics, San Diego, California, 2016, pp. 546–559. URL: <https://aclanthology.org/S16-1084>. doi:10.18653/v1/S16-1084.
- [41] J. Mitchell, M. Lapata, Composition in distributional models of semantics, *Cogn. Sci.* 34 (2010) 1388–1429. URL: <https://doi.org/10.1111/j.1551-6709.2010.01106.x>. doi:10.1111/j.1551-6709.2010.01106.x.
- [42] S. Reddy, D. McCarthy, S. Manandhar, An empirical study on compositionality in compound nouns, in: *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011*, Chiang Mai, Thailand, November 8-13, 2011, The Association for Computer Linguistics, 2011, pp. 210–218. URL: <https://aclanthology.org/I11-1024/>.
- [43] M. Farahmand, J. Henderson, Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model, in: *Proceedings of the 12th Workshop on Multiword Expressions*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 61–66. URL: <https://www.aclweb.org/anthology/W16-1809>. doi:10.18653/v1/W16-1809.

- [44] S. Cordeiro, A. Villavicencio, M. Idiart, C. Ramisch, Unsupervised compositionality prediction of nominal compounds, *Computational Linguistics* 45 (2019) 1–57. URL: <https://www.aclweb.org/anthology/J19-1001>. doi:10.1162/coli_a_00341.
- [45] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, *CoRR* abs/1310.4546 (2013). URL: <http://arxiv.org/abs/1310.4546>. arXiv:1310.4546.
- [46] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014*, pp. 1532–1543. URL: <https://doi.org/10.3115/v1/d14-1162>. doi:10.3115/v1/d14-1162.
- [47] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [48] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: <https://www.aclweb.org/anthology/N18-1202>. doi:10.18653/v1/N18-1202.
- [49] T. Schuster, O. Ram, R. Barzilay, A. Globerson, Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1599–1613. URL: <https://aclanthology.org/N19-1162>. doi:10.18653/v1/N19-1162.
- [50] V. Shwartz, I. Dagan, Still a pain in the neck: Evaluating text representations on lexical composition, *Transactions of the Association for Computational Linguistics* 7 (2019) 403–419. URL: <https://aclanthology.org/Q19-1027>. doi:10.1162/tac1_a_00277.
- [51] N. Nandakumar, T. Baldwin, B. Salehi, How Well Do Embedding Models Capture Non-compositionality? A View from Multiword Expressions, in: *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, Association for Computational Linguistics, Minneapolis, USA, 2019, pp. 27–34. URL: <https://www.aclweb.org/anthology/W19-2004>. doi:10.18653/v1/W19-2004.
- [52] M. King, P. Cook, Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of english verb-noun combinations, in: I. Gurevych, Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, Association for Computational Linguistics, 2018, pp. 345–350. URL: <https://www.aclweb.org/anthology/P18-2055/>. doi:10.18653/v1/P18-2055.
- [53] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Probing for id-

- iomaticity in vector space models, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3551–3564. URL: <https://aclanthology.org/2021.eacl-main.310>.
- [54] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels, 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), Association for Computational Linguistics, Online, 2021, pp. 2730–2741. URL: <https://aclanthology.org/2021.acl-long.212>.
- [55] H. Tayyar Madabushi, E. Gow-Smith, C. Scarton, A. Villavicencio, Astitchintransformers: Dataset and methods for the exploration of idiomaticity in pre-trained language models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing., Association for Computational Linguistics, 2021.