semantic similarity and sentence-level Natural Language Understanding. In particular, Multi-SimLex fills a gap in available resources for multilingual NLP and might help understand how lexical and compositional semantics interact if put alongside existing resources such as XNLI (Conneau et al. 2018b) for natural language inference or PAWS-X (Yang et al. 2019) for cross-lingual paraphrase identification. Finally, the Multi-SimLex annotation could turn out to be a unique source of evidence to study the effects of polysemy in human judgments on semantic similarity: for equivalent word pairs in multiple languages, are the similarity scores affected by how many senses the two words (or multi-word expressions) incorporate?

In light of the success of initiatives like Universal Dependencies for multilingual treebanks, we hope that making Multi-SimLex and its guidelines available will encourage other researchers to expand our current sample of languages. We particularly encourage creation and submission of comparable Multi-SimLex datasets for under-resourced and typologically diverse languages in future work. In particular, we have made a Multi-Simlex community website available to facilitate easy creation, gathering, dissemination, and use of annotated datasets: https://multisimlex.com/.

Acknowledgments

This work is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). Thierry Poibeau is partly supported by a PRAIRIE 3IA Institute fellowship ("Investissements d'avenir" program, reference ANR-19-P3IA-0001).

References

Adams, Oliver, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalová, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL-HLT*, pages 19–27.

Aldarmaki, Hanan and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Proceedings of NAACL-HLT*, pages 3906–3911.

Alvarez-Melis, David and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of EMNLP*, pages 1881–1890.

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the ACL*, 4:431–444.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, pages 451–462.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of AAAI*, pages 5012–5019.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

Artetxe, Mikel, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of CoNLL*, pages 282–291.

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL*, pages 86–90.

Baker, Simon, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of EMNLP*, pages 278–289.

Bapna, Ankur and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of EMNLP*, pages 1538–1548.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The