Multi-SimLex: A Large-Scale Evaluation of Multilingual and Cross-Lingual Lexical Semantic Similarity

https://multisimlex.com/

Ivan Vulić *♠
LTL, University of Cambridge

Simon Baker *♠
LTL, University of Cambridge

Edoardo Maria Ponti *♠
LTL, University of Cambridge

Ulla Petti *
LTL, University of Cambridge

Ira Leviant **
Faculty of Industrial Engineering and Management, Technion, IIT

Kelly Wing *
LTL, University of Cambridge

Olga Majewska *
LTL, University of Cambridge

Eden Bar **
Faculty of Industrial Engineering and Management, Technion, IIT

Matt Malone *
LTL, University of Cambridge

Thierry Poibeau †
LATTICE Lab, CNRS and ENS/PSL and Univ. Sorbonne nouvelle/USPC

Roi Reichart **
Faculty of Industrial Engineering and Management, Technion, IIT

Anna Korhonen *
LTL, University of Cambridge

We introduce Multi-SimLex, a large-scale lexical resource and evaluation benchmark covering datasets for 12 typologically diverse languages, including major languages (e.g., Mandarin Chinese, Spanish, Russian) as well as less-resourced ones (e.g., Welsh, Kiswahili). Each language dataset is annotated for the lexical relation of semantic similarity and contains 1,888 semantically aligned concept pairs, providing a representative coverage of word classes (nouns, verbs, adjectives, adverbs), frequency ranks, similarity intervals, lexical fields, and concreteness levels. Additionally, owing to the alignment of concepts across languages, we provide a suite of 66 cross-lingual semantic similarity datasets. Due to its extensive size and language coverage, Multi-SimLex provides entirely novel opportunities for experimental evaluation and analysis. On its monolingual

* ♠*Equal contribution*; English Faculty Building, 9 West Road Cambridge CB3 9DA, United Kingdom. E-mail: {iv250,sb895,ep490,om304,alk23}@cam.ac.uk
** Technion City, Haifa 3200003, Israel. E-mail: {ira.leviant,edenb}@campus.technion.ac.il, roiri@ie.technion.ac.il
† Rue Maurice Arnoux, 92120 Montrouge, France. E-mail: thierry.poibeau@ens.fr