Extending the key semantic domains method beyond English corpora: Wmatrix version 5

Paul Rayson, Lancaster University

The key semantic domains method (Rayson, 2008) implemented in Wmatrix (versions 1 to 4) extends the keywords approach which has been widely applied in corpus linguistics research. Key semantic domains facilitates the discovery of concepts and groups of words collected within semantic fields which are unusually frequent or infrequent compared to a reference corpus, and can exploit significance and effect size measures in the same way as the key words approach. Key semantic domains have proved useful in a number of different areas of linguistic research: literary characterisation (Balossi, 2014), language of psychopaths (Hancock et al., 2013), corpus-assisted discourse analysis of social work writing (Leedham et al., 2020), enhancing critical thinking in higher education (O'Halloran, 2020), and the construction of newsworthiness (Potts et al., 2015). However, one important drawback is that key semantic domains are currently restricted to one language only due to the inclusion of the CLAWS Part-of-Speech (POS) tagger (Garside and Smith, 1997) and the UCREL Semantic Analysis System (USAS) for English (Rayson et al., 2004). In recent years, semantic taggers for other languages have been developed (Piao et al., 2015; Piao et al., 2016) utilising freely available POS taggers and lemmatisers for new languages, and adapting a variety of methods ranging from bilingual dictionaries, parallel aligned corpora, machine translation, and crowdsourcing to bootstrap development of new semantic lexicons, and vector-based, pre-trained embeddings and machine learning methods to improve contextual disambiguation (Ezeani et al., 2019). Previously, a beta version of the Spanish semantic tagger has been incorporated into Wmatrix4. This poster will describe how the semantic taggers for further languages are being incorporated into Wmatrix5. Crucially, there is a need to support community crowdsourcing involvement for the extension and checking of the new semantic lexicons which are under varying stages of development to improve their coverage and accuracy.

Acknowledgements:

The new semantic lexicons are freely available under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (https://github.com/UCREL/Multilingual-USAS) and their application via semantic taggers in Wmatrix5 will continue to be free to use. A number of different international teams have been involved in the development, see http://ucrel.lancs.ac.uk/usas/ for details.

References:

Balossi, G. (2014) A Corpus Linguistic Approach to Literary Language and Characterization Virginia Woolf's The Waves. Benjamins.

Ezeani, I., Piao, S., Neale, S., Rayson, P., & Knight, D. (2019). Leveraging Pre-Trained Embeddings for Welsh Taggers. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019): Held at the 57th Annual Meeting of the Association for Computational Linguistics (pp. 270-280). Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-4332

Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121.

Hancock, J.T., Woodworth, M.T., and Porter, S. (2013) Hungry like the wolf: A word-pattern analysis of the language of psychopaths. Legal and Criminological Psychology. 18 (1), 102-114. http://dx.doi.org/10.1111/j.2044-8333.2011.02025.x

Leedham, M., Lillis, T. & Twiner, A. (2020). Exploring the core 'preoccupation' of social work writing: A corpus-assisted discourse study. Journal of Corpora and Discourse Studies. 3. Pp.1-26. https://jcads.cardiffuniversitypress.org/articles/abstract/26/

O'Halloran, K. (2020). A posthumanist pedagogy using digital text analysis to enhance critical thinking in higher education, Digital Scholarship in the Humanities, 35 (4) 845–880. https://doi.org/10.1093/llc/fqz060

Potts, A., Bednarek, M., and Caple, H. (2015). How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. Discourse & Communication, 9 (2) 149 - 172. https://doi.org/10.1177/1750481314568548

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A. and Rayson, P. (2015). Development of the multilingual semantic annotation system. In proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015), Denver, Colorado, United States, pp. 1268-1274. https://www.aclweb.org/anthology/N15-1137/

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R., Knight, D., Kren, M., Löfberg, L., Nawab, R.M.A., Shafi, J., Teh, P.L. and Mudraya, O. (2016) Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. In proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016), Portoroz, Slovenia, pp. 2614-2619. https://www.aclweb.org/anthology/L16-1416/

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.

Rayson, P. (2008). From key words to key semantic domains. International Journal of Corpus Linguistics. 13 (4) 519-549. https://doi.org/10.1075/ijcl.13.4.06ray