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

Paul Rayson

School of Computing and Communications

Data Science | Lancaster University

Abstract

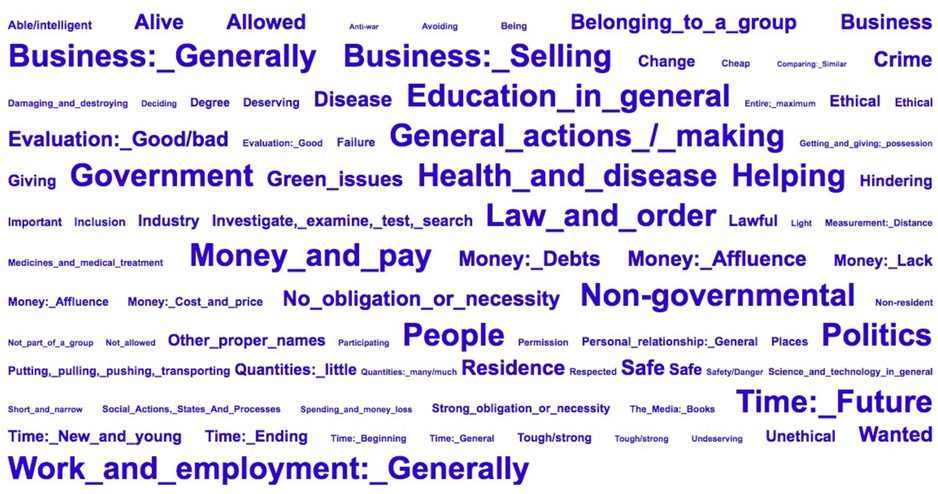
- The key semantic domains method implemented in Wmatrix (versions 1 to 4) extends the keywords approach which has been widely applied in corpus linguistics research.
- However, one important drawback is that key semantic domains are currently restricted to one language only due to the inclusion of the CLAWS Part-of-Speech (POS) tagger and the UCREL Semantic Analysis System (USAS) for English.
- In recent years, semantic taggers for other languages have been developed utilising freely available POS taggers and lemmatisers for new languages.
- This poster describes 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.
- This work will enable key semantic domains for monolingual analysis beyond English corpus, but also facilitate crosslingual comparisons.

Key Semantic Domains

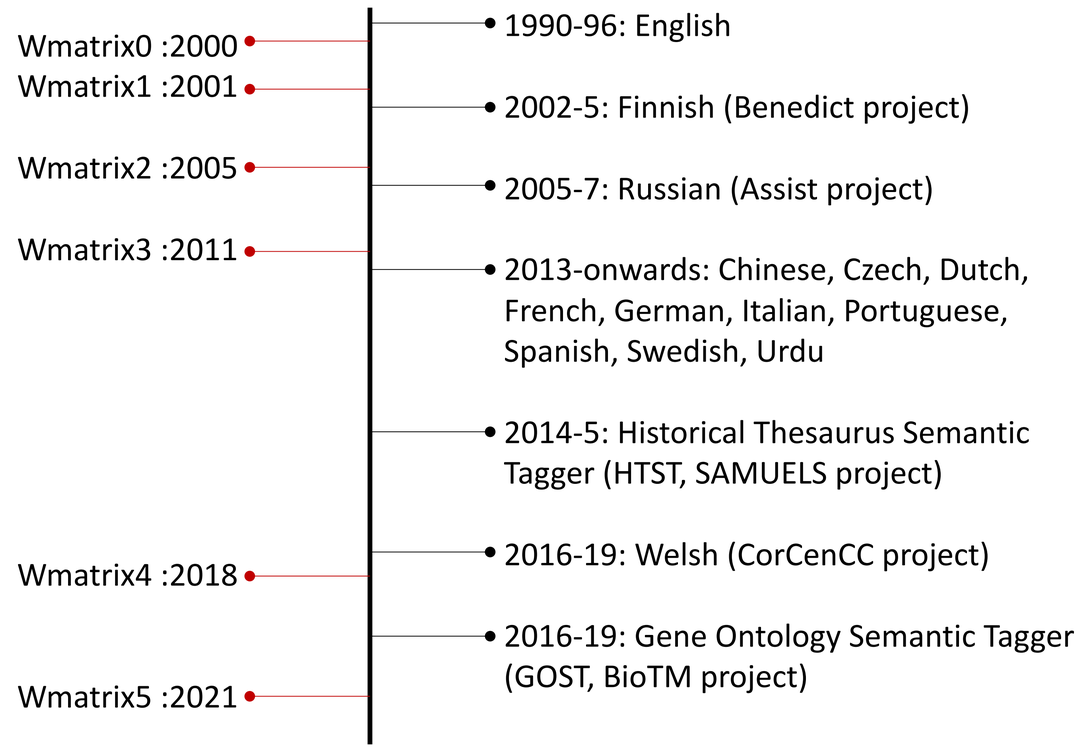
- Key semantic domains facilitates the discovery of concepts and groups of words collected within semantic fields which are unusually frequent or infrequent compared to a reference corpus.
- Key semantic domains have proved useful in a number of different areas of linguistic research: literary characterisation (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).

- Mini case study: Wmatrix corpus analysis of UK General Election Manifestos 2017 http://ucrel.lancs.ac.uk/wmatrix/ukmanifestos2017/
- Figure 1: Labour Key Word Cloud

- Figure 2: Labour Key Semantic Tag Cloud



Wmatrix and Semantic Tagger Development Timelines



The Future ... New Semantic Taggers Currently Planned

• Arabic	Nouran Khallaf (Leeds, UK), Elvis de Souza (PUC-Rio, Brazil), Mahmoud El-Haj (Lancaster, UK)
• Indonesian	Prihantoro (Lancaster, UK)
• Korean	Se-Eun Jhang (KMOU, Corpus Linguistics Research Association under the auspices of The Korean Association of Language Sciences, Korea)
• Persian	Mehrdad Vasheghani Farahani (Leipzig, Germany)

Requirements For Adding New Languages

- Reference corpora for each language.
- Corpus indexing system: MatrixDB and/or LexiDB (Coole, 2021).
- Personal (private) lexicons versus system (public) lexicons.
- Support for crowdsourcing and checking of new semantic lexicons.
- Manually checked gold standard data.
- Existing POS tagger and lemmatiser.
- Bootstrapping methods to create initial semantic lexicon (bilingual dictionary, parallel corpus, MT, NER, ML/DL, vector-based).

Wmatrix https://ucrel-wmatrix5.lancaster.ac.uk/

Open Access

See http://ucrel.lancs.ac.uk/wmatrix/ for details on which components of Wmatrix are already open source, freely available, and/or accessible via REST APIs. More elements will be released soon. Accounts on Wmatrix5 are freely available for Lancaster alumni, UK government, and academic research in countries on the OECD DAC list of ODA recipients (https://www.oecd.org/).

Acknowledgements

Wmatrix was initially developed within the REVERE project (REVerse Engineering of Requirements) funded by the EPSRC, 1998-2001. Ongoing maintenance of taggers (e.g. Linux porting work by Stephen Wattam), development of new components (e.g. dictionary updates by Sheryl Prentice) are funded by user licence fees. Semantic taggers for new languages have been developed by Scott Piao also funded by UCREL. The new semantic lexicons are freely available under CC-BY-NC-SA-4.0 licence (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. Metaphor extensions have been developed in the MELC project funded by the ESRC. The Historical Thesaurus Semantic Tagger (HTST) was developed in the SAMUELS project funded by the AHRC in conjunction with the ESRC, with MMU, Glasgow and Huddersfield Universities, and OUP. BioTM project: joint work with Jo Knight, Mahmoud El-Haj, Scott Piao, funded by the Wellcome Trust.