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



Paul Rayson **School of Computing and Communications**

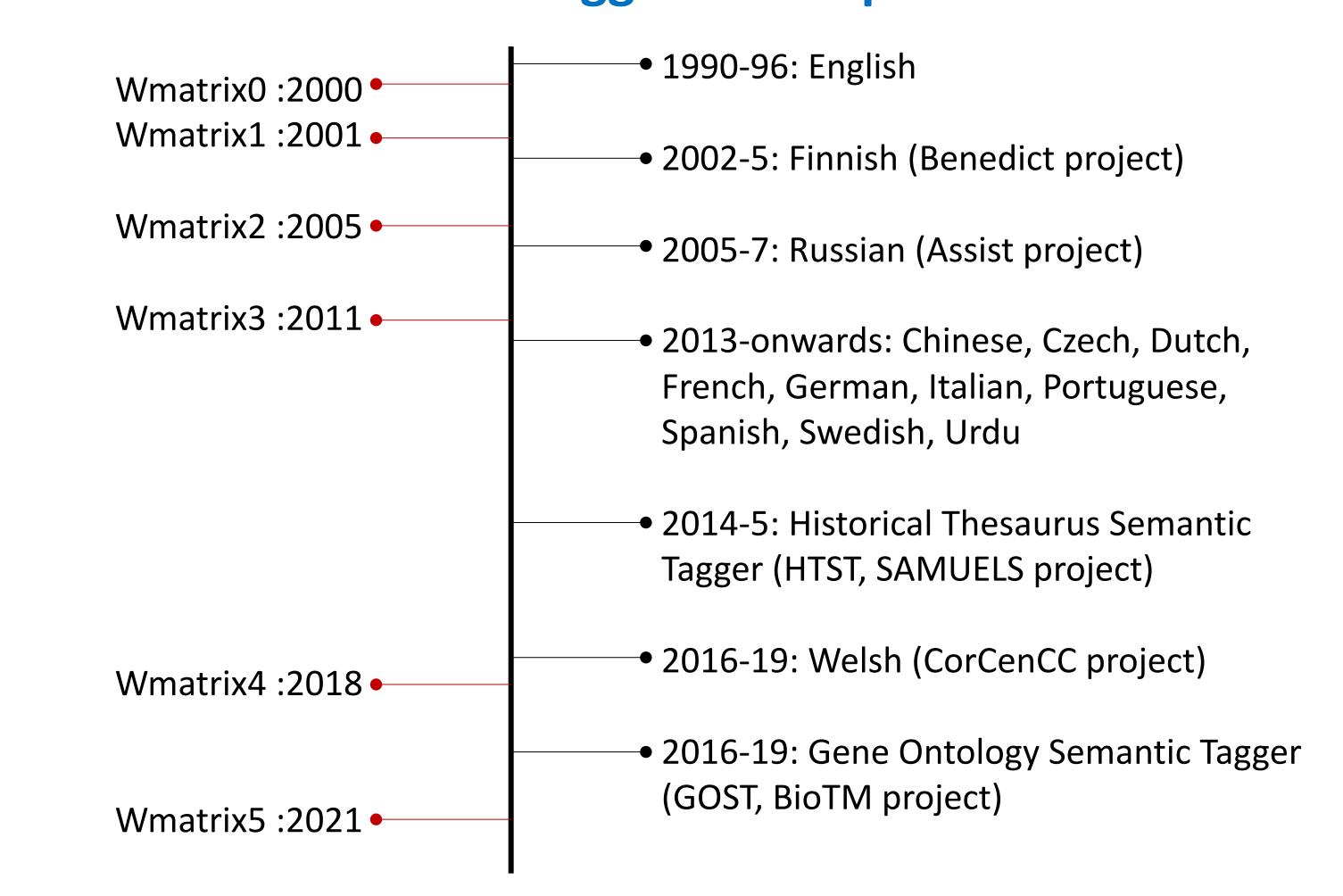




Abstract

- 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.

Wmatrix and Semantic Tagger Development Timelines



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

01 09 access across affordable and brexit Britain build businesses care Chapter communities con conservative Conservatives consult country crime cuts deliver devolved disabilities discrimination e economy education employers energy ensure ensuring equality EU fair funding future global government guarantee health healthcare held_back homes housing ict improve including infrastructure introduce invest investment jobs justice justice labour legislate LGBT LL local mental_health n national negotiating new NHS our ownership patients pay people poverty prisons protect protections public reduce reinstate review richer rights s s schools scrap sector security Services social_security standards strategy support sustainable to trade_unions We Will work workers workplace

Figure 2: Labour Key Semantic Tag Cloud

Business:_Generally Business:_Selling Change Cheap Comparing:_Similar Crime Damaging_and_destroying Deciding Degree Deserving Disease Education_in_general Entire;_maximum Ethical Ethical Evaluation:_Good/bad Evaluation:_Good Failure General_actions_/_making Getting_and_giving;_possession Giving Government Green_issues Health_and_disease Helping Hindering Important Inclusion Industry Investigate,_examine,_test,_search Law_and_order Lawful Light Measurement:_Distance Medicines_and_medical_treatment Money_and_pay Money:_Debts Money:_Affluence Money:_Lack Money:_Affluence Money:_Cost_and_price No_obligation_or_necessity Non-governmental Non-resident Not_part_of_a_group Not_allowed Other_proper_names Participating People Permission Personal_relationship:_General Places Politics Putting,_pulling,_pushing,_transporting Quantities:_little Quantities:_many/much Residence Respected Safe Safe Safe Safety/Danger Science_and_technology_in_general Short_and_narrow Social_Actions,_States_And_Processes Spending_and_money_loss Strong_obligation_or_necessity The_Media:_Books Time: Future Time:_New_and_young Time:_Ending Time:_Beginning Time:_General Tough/strong Tough/strong Undeserving Work_and_employment:_Generally

The Future ... New Semantic Taggers Currently Planned

Nouran Khallaf (Leeds, UK), Elvis de Souza (PUC-Rio, Arabic Brazil), Mahmoud El-Haj (Lancaster, UK)

Prihantoro (Lancaster, UK) Indonesian

Se-Eun Jhang (KMOU, Corpus Linguistics Research Korean

Association under the auspices of The Korean Association of Language Sciences, Korea)

Mehrdad Vasheghani Farahani (Leipzig, Germany) Persian

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).

Шmatrix

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.