LDK 2021 - 3rd Conference on Language, Data and Knowledge, 1-4 September, Zaragoza, Spain

HISTORIAE, HIstory of Socio-culTural transfORmation as linguistIc dAta sciEnce. A Humanities Use Case

Florentina Armaselu, florentina.armaselu@uni.lu, University of Luxembourg, Luxembourg; **Elena-Simona Apostol**, elena.apostol@upb.ro, University Politehnica of Bucharest, Romania; **Anas Fahad Khan**, fahad.khan@ilc.cnr.it, Istituto di Linguistica Computazionale «A. Zampolli», Italy; **Chaya Liebeskind**, liebchaya@gmail.com, Jerusalem College of Technology, Israel; **Barbara McGillivray**, bmcgillivray@turing.ac.uk, The Alan Turing Institute, United Kingdom; **Ciprian-Octavian Truică**, ciprian.truica@upb.ro, University Politehnica of Bucharest, Romania; **Giedrė Valūnaitė Oleškevičienė**, gvalunaite@mruni.eu, Mykolas Romeris University, Lithuania

Context

Use case initiated within the COST Action **Nexus Linguarum, European network for Web centred linguistic data science** (CA18209), 2019-2023. https://nexuslinguarum.eu/.

Goals

Create a **comparative methodological framework** for tracing the "histories" or evolution of concepts in different languages and humanities fields (history, literature, philosophy, religion, etc.) and generate a sample of **multilingual LLOD ontologies** to represent semantic change and corresponding **explanations**, by using NLP and Semantic Web technologies.

Concepts

Domain: socio-cultural transformation.

Semantic fields: geo-political and cultural entities (Europe, West, East, etc.), education, sciences, technology and innovations, social and societal processes (migration, urbanisation, modernisation, globalisation), state and citizenship, beliefs, values and attitudes (e.g. religion, democracy, political participation), economy, health and well-being, everyday life, family and social relations, time and collective memory, work and leisure, customs and traditions, literature and philosophy.

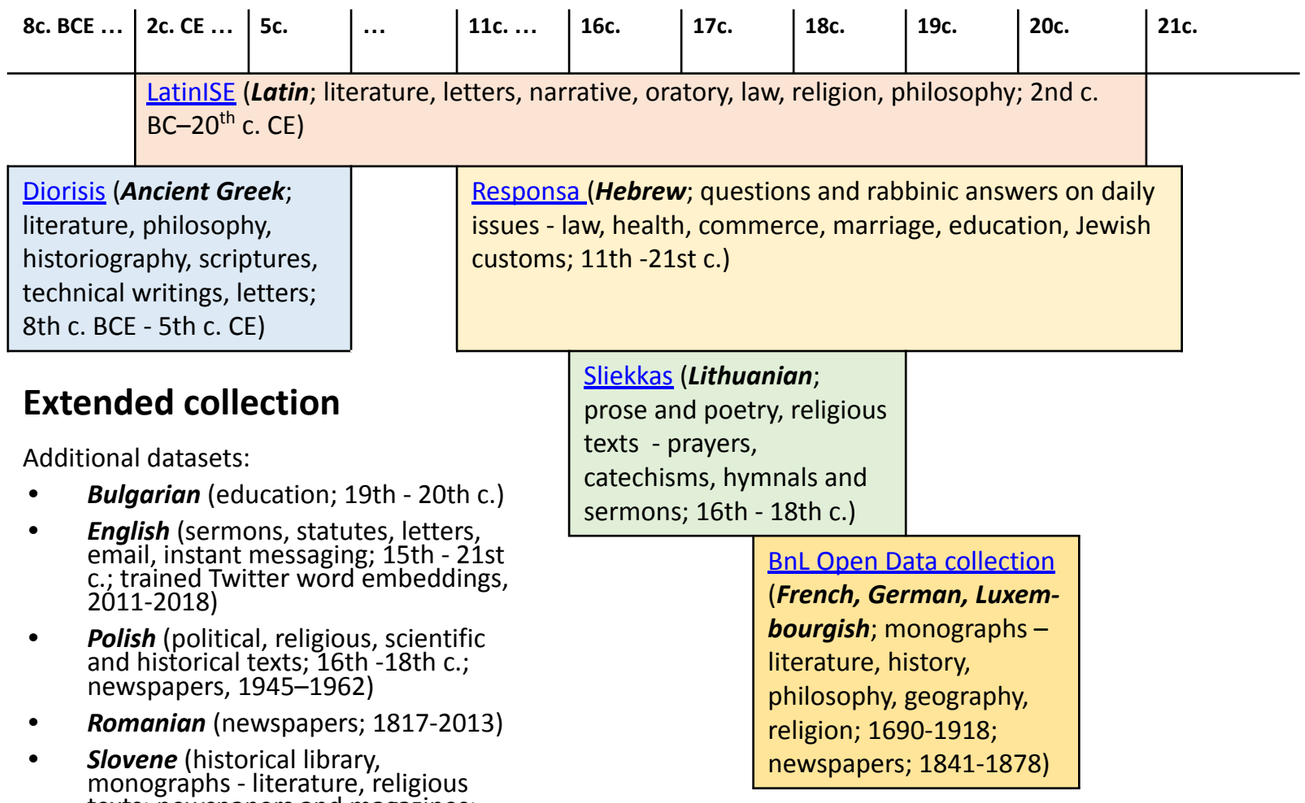
Questions

Can the applied methodology inform us about the **interrelation** between **linguistic**, **social** and **cultural innovation** over time, and the socio-cultural roots of innovation?

What may be learned about the combination of **human** and **machine agency** in the process of **construction** and **dissemination** of knowledge through NLP and Semantic Web technologies, and of **explaining** the underlying mechanisms?

Datasets, languages, content, time span

Core collection

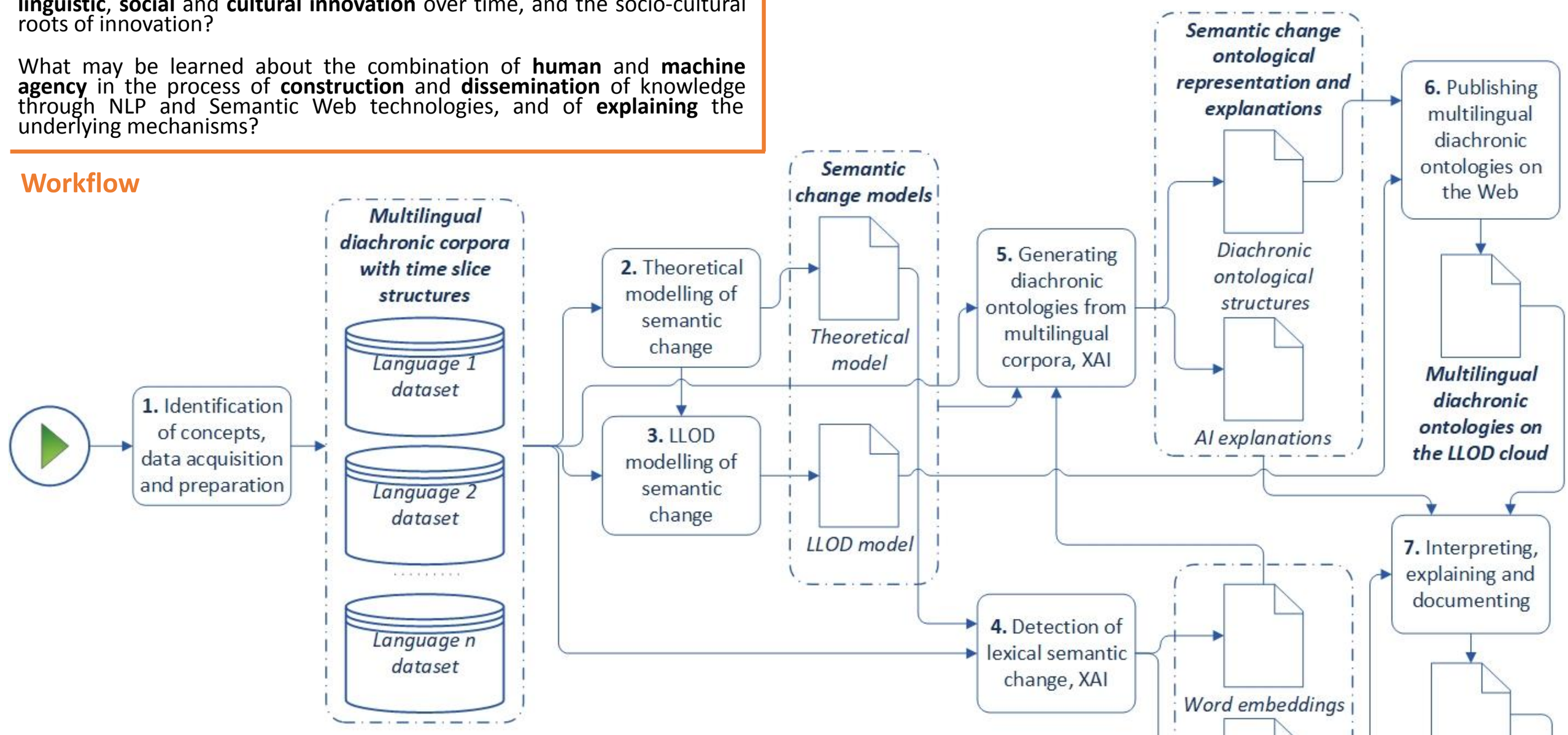


Extended collection

Additional datasets:

- ***Bulgarian*** (education; 19th - 20th c.)
- ***English*** (sermons, statutes, letters, email, instant messaging; 15th - 21st c.; trained Twitter word embeddings, 2011-2018)
- ***Polish*** (political, religious, scientific and historical texts; 16th -18th c.; newspapers, 1945–1962)
- ***Romanian*** (newspapers; 1817-2013)
- ***Slovene*** (historical library, monographs - literature, religious texts; newspapers and magazines; 1584-1919)
- ***Multilingual*** (Wikimedia downloads and archives; 2001-2021; parliamentary discourse; 2001 – 2012)

Workflow



Methodological starting points

Data preparation: conversions (XML to TXT), metadata extraction (language, date, genre), structuring by time slice (year, decade, century).

Theoretical modelling of semantic change: *semasiological* (new meanings) vs. *onomasiological* (new lexical items) innovation mechanisms [1]; concept *core* and *margin* variability [2]; *intension*, *extension* and *label* that define the meaning of a concept, distance measures [3].

LLOD modelling of semantic change: extensions of the OntoLex-Lemon model [4, 5] and the Lexical Markup Framework (LMF) to represent diachronic information [6].

Detection of lexical semantic change: word embeddings enriched with temporal-spatial information [7], SemEval-2020 task 1 (unsupervised lexical semantic change detection) [8], transformer-based semantic change detection [9]. https://radimrehurek.com/gensim/. https://github.com/huggingface/transformers.

Generating and publishing LLOD diachronic ontologies: *ontology learning layer cake* model (*terms, synonyms, concepts, concept hierarchies, relations* and *rules*) [10], Text2Onto [11], word2vec for ontology learning [12], diachronic ontologies building [13], transformation pipelines [14]. https://code.google.com/archive/p/text2onto/. https://github.com/Pret-a-LLOD/Fintan.

Explainable AI (XAI): four principles of explainable AI systems: *explanation* (the system is "capable of providing an explanation"), *meaningfulness* (the recipient "understands the system's explanations"), *explanation accuracy* (the explanation "may or may not accurately describe how the system came to its conclusion"), *knowledge limits* (the systems identify cases when they "were not designed or approved to operate", or "their answers are not reliable") [15].

Next steps

Hypothesis testing (can these various types of methods and tools be integrated into a coherent pipeline?) and **implementation** of the **workflow**.

References

[1] D. Geeraerts. *Theories of lexical semantics*. Oxford University Press, 2010. **[2]** J-M. Kuukkanen. "Making Sense of Conceptual Change." *History and Theory*, 47(3), 2008. **[3]** S. Wang, S. Schlobach, M. Klein. "Concept Drift and How to Identify It." *Journal of Web Semantics, First Look*, September 2011. **[4]** J.P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, Ph. Cimiano. "The OntoLex-Lemon Model: Development and Applications." *Electronic Lexicography in the 21st Century*. September 2017. **[5]** A.F. Khan. "Towards the Representation of Etymological Data on the Semantic Web." *Information*, 9(12), November 2018. **[6]** L. Romary, M. Khemakhem, F. Khan, J. Bowers, N. Calzolari, M. George, M. Pet, P. Bański. "LMF Reloaded." arXiv:1906.02136, 2019. **[7]** H. Gong, S. Bhat, P. Viswanath. "Enriching Word Embeddings with Temporal and Spatial Information." *The 24th Conference on Computational Natural Language Learning*, 2020. **[8]** D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi. "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection." *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020. **[9]** J. Rodina, Y. Trofimova, A. Kutuzov, E. Artemova. "ELMo and BERT in Semantic Change Detection for Russian." *CoRR*, 2020. arXiv:2010.03481. **[10]** P. Buitelaar, Ph. Cimiano, B. Magnini. "Ontology Learning from Text: An Overview." *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123, IOS Press, 2005. **[11]** Ph. Cimiano, J. Volker. "Text2Onto. A Framework for Ontology Learning and Data-driven Change Discovery." *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems*, 2005. **[12]** G. Wohlgenannt, F. Minic. "Using word2vec to Build a Simple Ontology Learning System." *International Semantic Web Conference*, 2016. **[13]** S. He, X. Zou, L. Xiao, J. Hu. "Construction of Diachronic Ontologies from People's Daily of Fifty Years." *LREC*, 2014. **[14]** C. Fäth, C. Chiarcos, B. Ebbrecht, M. Ionov. "Fintan - Flexible, Integrated Transformation and Annotation eNgineering." *The 12th Conference on Language Resources and Evaluation*, 2020. **[15]** P.J. Phillips, C.A. Hahn, P.C. Fontana, D.A. Broniatowski, M.A. Przybocki. "Four Principles of Explainable Artificial Intelligence." National Institute of Standards and Technology, U.S. Department of Commerce, August 2020.