University of Luxembourg

Multilingual. Personalised. Connected.

UNIVERSITÉ DU LUXEMBOURG

Bridging NLP and LLOD: Humanities Approaches to Semantic Change. Part 1

Florentina Armaselu, florentina.armaselu@uni.lu

KBR Digital Heritage Series, 8 December 2022

C²DH
LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

uni.lu
UNIVERSITÉ DU LUXEMBOURG

Summary

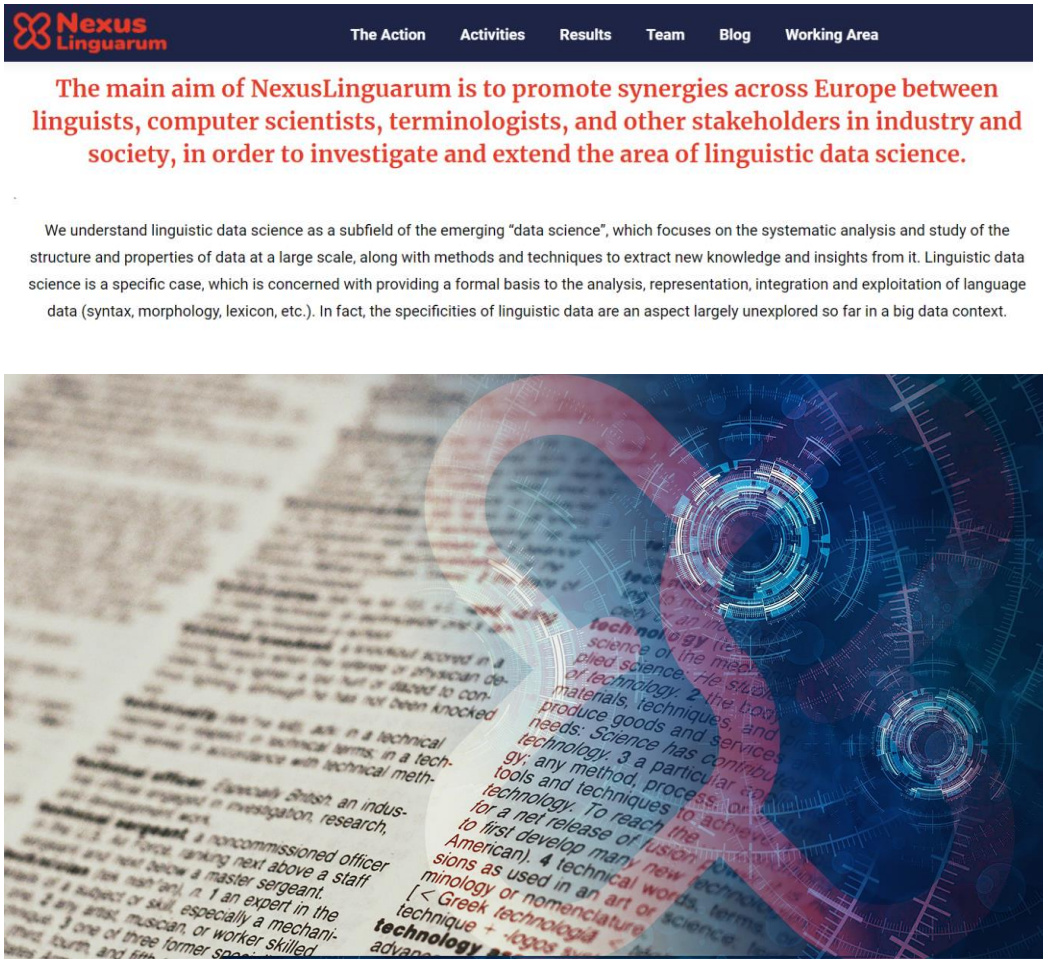
C²DH
LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

uni.lu
UNIVERSITÉ DU LUXEMBOURG

Part 1

Research background

1. About the project. *NexusLinguarum. Humanities use case. Overview*



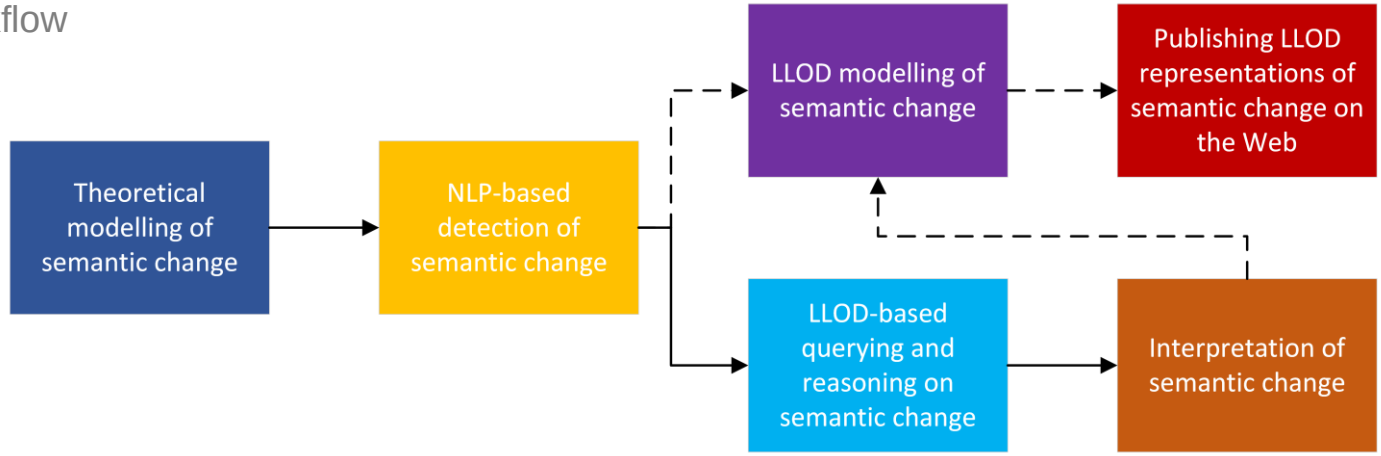
- *Nexus Linguarum* - European network for Web-centred linguistic data science (https://nexuslinguarum.eu/)
  - COST Action, CA18209 (2019-2024)
- Use case in the Humanities (UC4.2.1, working group 04) objectives
  - trace the evolution of (parallel) concepts in a collection of multilingual, diachronic corpora;
  - combine natural language processing (NLP) and linguistic linked open data (LLOD) to detect and model semantic change;
  - publish a sample of diachronic ontologies in the LLOD cloud.

1. About the project. *NexusLinguarum. Humanities use case. Research questions and workflow*

- Semantic change
  - change in meaning, either of a lexical unit (word or expression) or
  - of a concept (a complex knowledge structure that can encompass one or more lexical units as well as relations among them and with other concepts).

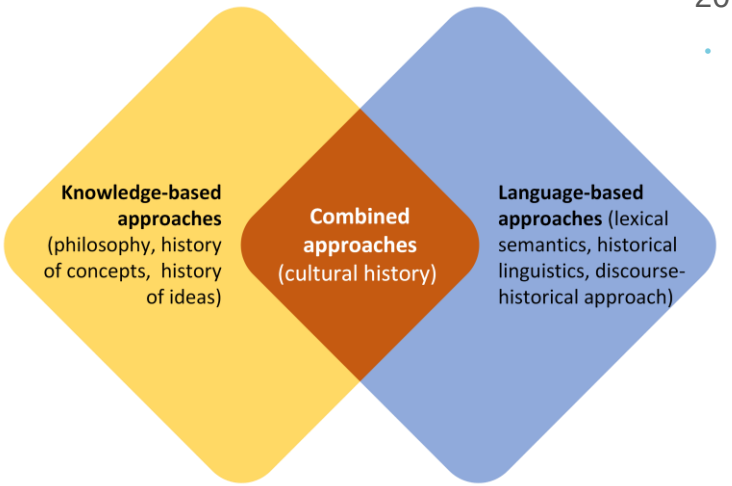
- Research questions
  - What are the mechanisms that determine linguistic innovation?
  - How are these mechanisms related to the reality?
  - Are these mechanisms language- and culture-specific or do they encompass universal aspects, applying to all languages?
  - Is it possible to detect, represent and reason about semantic change through NLP methods and linked data formalisms such as LLOD? What types of resources are needed to attain this goal?

- Workflow



2. Semantic change. *Theoretical frameworks*

- History of concepts (*Begriffs-geschichte*) (Koselleck 1994)
  - Relationship **concepts** – **reality** over a period of time:
    1) concepts and related reality – **stable**;
    2) concepts and reality **change** at the **same time**;
    3) **concept change** without a change in the reality;
    4) **reality changes**, concepts remain the same.
  - Concepts -> multi-layered **temporal structure** -> ties with:
    - past events and experiences;
    - present reality;
    - expectations for the future.
  - **Sources** -> temporal structure:
    - **instant** use (newspapers, letters, speeches);
    - gradual **development** (lexicons, dictionaries, encyclopaedias, handbooks);
    - **unchanging** forms (classical texts, timeless values).



- Lexical semantics (Geeraerts 2010):
  - **Mechanisms** of semantic change:
    1. **semasiological** (meaning-related) -> semasiological innovations endowing existing words with new meanings;
    2. **onomasiological** (naming-related) -> onomasiological innovations coupling "concepts to words in a way that is not yet part of the lexical inventory of the language" (p. 26).

- Cultural history:
  - **Language** (Richter 1994: 125, 126):
    - understood as an "agent and an indicator of structural changes",
    - that "both shaped and registered the process of change".

➢ *How to capture through digital means the dynamic interaction between conceptual and socio-cultural changes?*

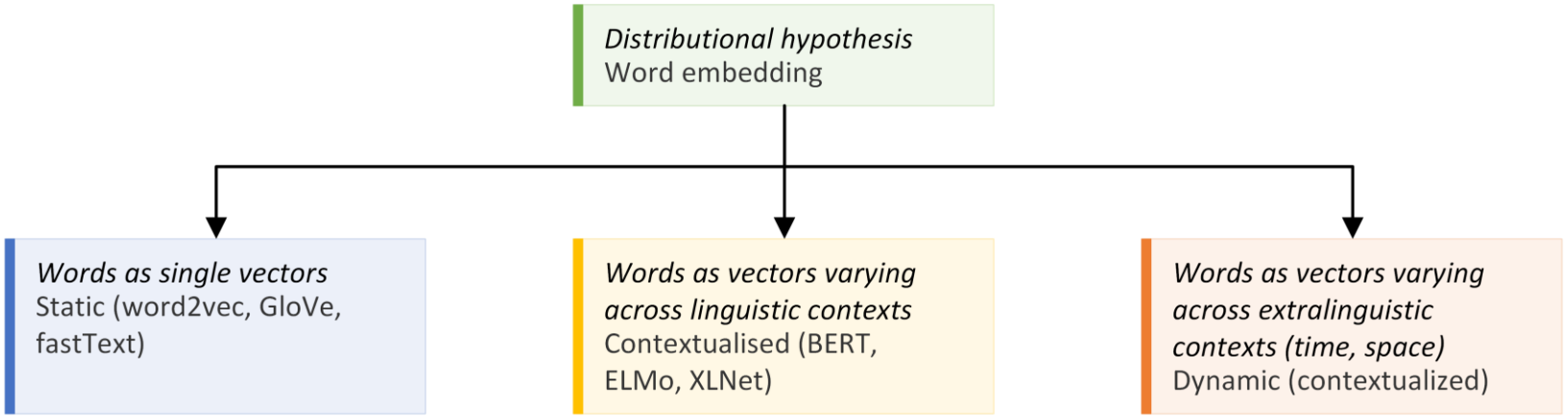
- Use of digital sources: **corpora** as a reflection of reality and **dictionaries** as normative resources.

2. Semantic change. *NLP approaches. Word embedding*

C²DH LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY
UNIVERSITÉ DU LUXEMBOURG

- Distributional hypothesis
  - words such as "*oculist* and *eye-doctor* […] occur in almost the **same environments**." (Harris 1954: 156)
  - **distributional semantics** -> quantifying "semantic similarities between linguistic items according to their distributional properties in large text corpora". (Goldberg, 2017: 118)
  - **word embedding** -> "learning representations of the meaning of words, called embeddings, directly from their distributions in texts". (Jurafsky and Martin, 2021: ch.6: 1)

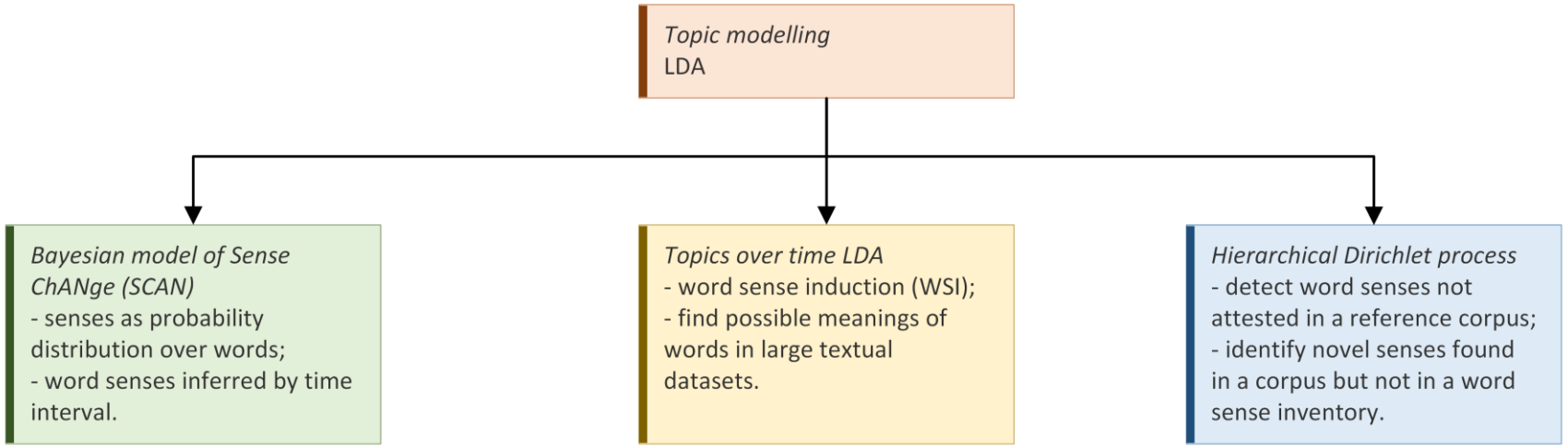
- Word embedding techniques used in "NLP **tasks** involving **semantic variability**" (Hofmann et al. 2021)



2. Semantic change. *NLP approaches. Topic modelling*

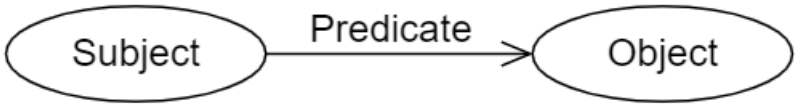
- Latent Dirichlet allocation (LDA) (Blei et al. 2003)
  - probabilistic technique -> representing each document in a corpus as a distribution over topics and each topic as a distribution over words.

- LDA as an element of **comparison** or as a basis for further **extensions** considering the temporal dimension in word meaning evolution (Armaselu et al. 2022a: 1065-1066)

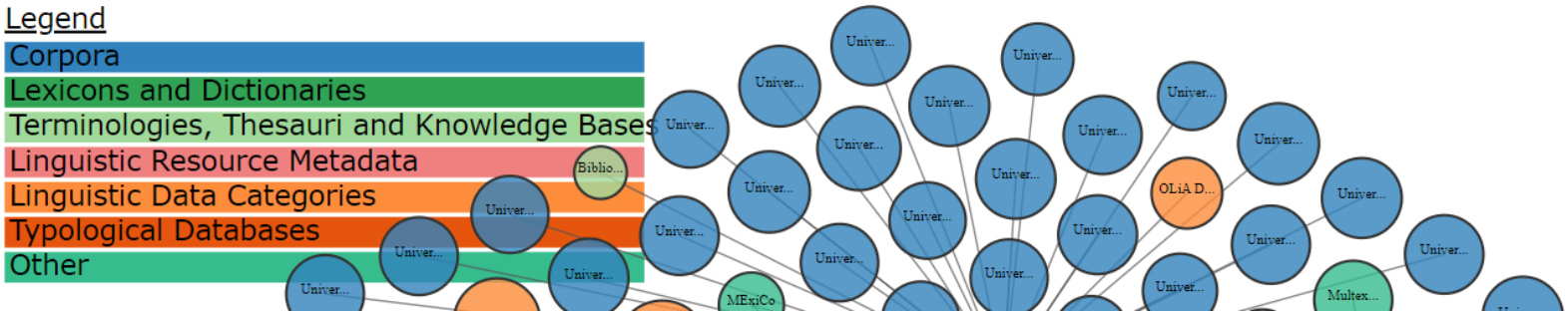


2. Semantic change. *LLOD formalisms (1)*

- Linguistic Linked Open Data (LLOD) (https://linguistic-lod.org/)
  - "movement about **publishing data for linguistics and natural language processing**" using "Web standards such as HTML, RDF or JSON-LD";
  - Resource Description Framework (RDF) -> framework for representing information on the Web as a set of **triples**: *subject, predicate* and *object* called an RDF graph, that can be visualised as a node and directed-arc diagram.



  - Linguistic Linked Open Data **cloud** diagram (Chiarcos et al. 2012)



2. Semantic change. *LLOD formalisms (2)*

- LLOD models
  - OntoLex-Lemon (McCrae 2017) -> model for **publishing lexicons** as linked data:
    - **LexicalEntry**: “head word” of a lexicon entry;
    - **Form**: written/spoken representation (morphological expressions);
    - **LexicalSense**: word sense;
    - **LexicalConcept**: equivalent lexical senses (synonyms).
  - Frequency, Attestation and Corpus Information module (OntoLex-FrAC) (Chiarcos et al. 2022) -> model integrating **lexical resources** and **information derived from corpora**:
    - **Observable**: observed entity (e.g., in a corpus) (e.g., LexicalEntry);
    - **Observation**: information based on or created from a corpus (e.g., *Attestation* – link to a lexicographical or corpus source; *Frequency* – absolute, relative frequency in a corpus; *Embeddings* – vector size, values).
  - Encoding of **etymological** information (Khan. 2018), extension of OntoLex-Lemon
    - **Etymology, Etymon**, **EtymologicalLink** - representing etymological entities and relationships.
- Other models (RDF limitation, not possible to add a temporal parameter to binary properties)
  - OWL-based, ***perdurantist*** approach - modelling entities as having temporal parts (Welty 2006).
  - OWL-Time - allowing to encode **temporal aspects** in RDF (Hobbs and Pan 2006).

C²DH
LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

uni.lu
UNIVERSITÉ DU LUXEMBOURG

Part 2

Nexus Linguarum humanities use case. Ongoing experiments

(currently unavailable, considered for publication)

5. Conclusion and future work

- Combine **NLP** and **LLOD corpus-** and **dictionary-based resources** to enable reasoning in the task of semantic change detection and representation within a multi-lingual context
  - integrate **attestation** from **dictionaries** (dates, citations) with **corpus information** (context examples, frequency, neighbours, embeddings, similarity or other types of measures);
  - capture the **interconnections** between **language** (attested by dictionaries) and **reality** change (as expressed in corpora);
  - enable **intra-** and **cross-lingual** inferences through **etymology** and **translation** relations;
  - record new **lexical forms/meanings**, when corpus evidence is identified;
  - detect **emerging concepts** (clusters of meaning), if possible, **before** having been **lexicalised**;
  - include persistent links to **digital facsimiles** (when available).

- Starting points
  - Combine diachronic word embedding results with LLOD formalisms, such as OntoLex-FrAC (Chiarcos et al. 2022).
  - Consider as an initial dictionary framework DBnary (Sérasset, 2015) – a periodically updated version of Wiktionary in linked data format.

References (1)

- Armaselu, Florentina. Apostol, Elena-Simona. Khan, Anas Fahad. Liebeskind, Chaya. McGillivray, Barbara. Truică, Ciprian-Octavian. Utka, Andrius. Valūnaitė Oleškevičienė, Giedrė. Van Erp, Marieke. 2022a. "LL(O)D and NLP perspectives on semantic change for humanities research." In *Semantic Web Journal*, Julia Bosque-Gil, Philipp Cimiano and Milan Dojchinovski (eds.), volume 13, IOS Press, pp. 1051–1080,. 2022a DOI: 10.3233/SW-222848. https://content.iospress.com/articles/semantic-web/sw222848.
- Armaselu, Florentina. Apostol, Elena-Simona. Chiarcos, Christian. Khan, Anas Fahad. Liebeskind, Chaya. McGillivray, Barbara. Truică, Ciprian-Octavian. Valūnaitė Oleškevičienė, Giedrė. 2022b. "Tracing Semantic Change with Multilingual LLOD and Diachronic Word Embeddings." Presentation at the LLODREAM2022 Conference, *LLOD Approaches for Language Data Research And Management*, Mykolas Romeris University, Vilnius, Lithuania, 21-22 September 2022. https://repository.mruni.eu/handle/007/18678.
- Blei, David M. Ng, Andrew Y. Jordan, Michael I. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, John Lafferty (Ed.), 993–1022.
- Bojanowski, Piotr. Grave, Edouard. Joulin, Armand. Mikolov, Tomas. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5, 135–146. DOI: 10.1162/tacl_a_00051.
- Chiarcos, Christian. Hellmann, Sebastian. Nordhoff, Sebastian. 2012. Linking linguistic resources: Examples from the Open Linguistics Working Group, In: Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), Linked Data in Linguistics. Representing Language Data and Metadata, Springer, Heidelberg, p. 201-216.
- Chiarcos, Christian; Apostol, Elena-Simona; Kabashi, Besim; Truică, Ciprian-Octavian. 2022. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. *Proceedings of the 29th International Conference on Computational Linguistics*, 4018–4027.
- Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford; New York: Oxford University Press. ISBN: 978-0-19-870031-9.
- Goldberg, Yoav. *Neural network methods for natural language processing*, University of Toronto, 2017.
- Harris, Zellig S. 1954. "Distributional Structure." *WORD* 10 (2–3): 146–62. doi:10.1080/00437956.1954.11659520.

References (2)

- Hofmann, Valentin. Pierrehumbert, Janet. Schütze, Hinrich. 2021. “Dynamic Contextualized Word Embeddings.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 6970–84. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.542.
- Hobbs, Jerry R. Pan, Feng. 2006. Time ontology in OWL, p. 133, W3C working draft 27.
- Jurafsky, Dan. Martin, James H. Speech and Language Processing (3rd ed. draft), chapter 6 December 2021. http://web.stanford.edu/~jurafsky/slp3/.
- Khan, Fahad. 2018.Towards the representation of etymological and diachronic lexical data on the semantic web, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, European Language Resources Association (ELRA), Miyazaki, Japan.
- Koselleck, Reinhart. 1994. Some Reflections on the Temporal Structure of Conceptual Change. *Main Trends in Cultural History.* Ten Essays. Eds. Melching, Willem; Velema, Wyger. Amsterdam - Atlanta, GA: Editions Rodopi, 7–16.
- McCrae, John P, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. 2017. “The OntoLex-Lemon Model: Development and Applications.” In Proceedings of ELex 2017 Conference. http://john.mccr.ae/papers/mccrae2017ontolex.pdf.
- Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs].
- Rehurek, Radim; Sojka, Petr. 2010. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 45–50. DOI: 10.1016/j.laa.2005.07.021.
- Richter, Melvin. 1994. Begriffsgeschichte in Theory and Practice: Reconstructing the History of Political Concepts and Languages. *Main Trends in Cultural History*. Ten Essays. Eds. Melching, Willem; Velema, Wyger. Amsterdam - Atlanta, GA: Editions Rodopi, 121–149.
- Sérasset, Gilles. 2015. “DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF.” Edited by Sebastian Hellmann, Steven Moran, Martin Brümmer, and John McCrae. Semantic Web 6 (4): 355–61. https://doi.org/10.3233/SW-140147.
- Welty, Christopher. Fikes, Richard. Makarios, Selene. 2006. A reusable ontology for fluents in OWL, in: *FOIS*, Vol. 150, pp. 226–236.

Stay connected with us!

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

@C2DH_LU

https://www.facebook.com/c2dh.lu/

www.c2dh.uni.lu