LT for DH: Language Technology Resources and Tools for Digital Humanities

# Tools and Instruments for Building and Querying Diachronic Computational Lexica

*Anas Fahad Khan  Andrea Bellandi  Monica Monachini*
Istituto di Linguistica Computazionale "A . Zampolli", CNR Pisa, Italy
Dipartimento di Studi Umanistici, Università di Ca' Foscari
name.surname@ilc.cnr.it

## Context and Aims

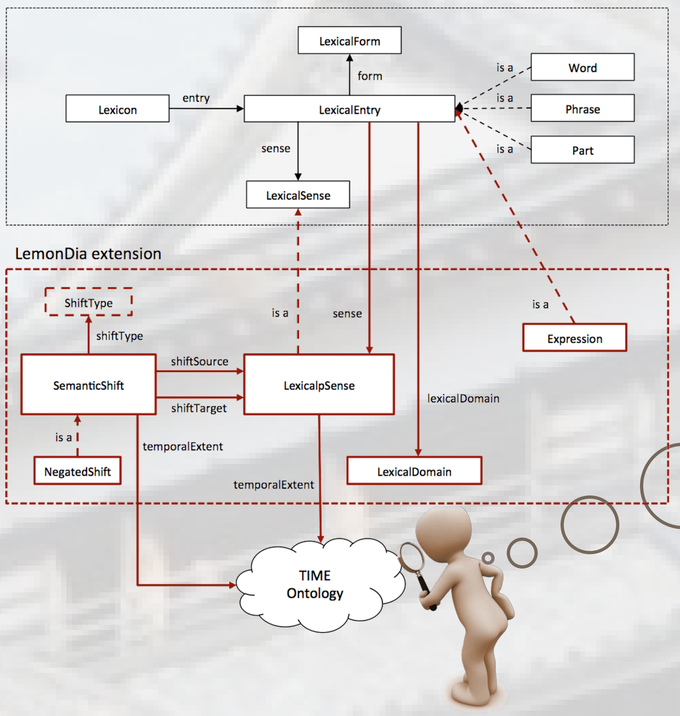
- It's important for "classical" languages, such as Ancient Greek, along with languages with a sizeable corpus of historical texts, that we have a flexible and accessible way of representing diachronic lexical information in a digital format.
- We focus on the represenation of semantic meaning shift.
- We need a formalism that's expressive enough, but that is still tractable.

## Linguistic Linked Open Data

- We chose Linked Data(LD) as our formalism because of how easy it is to publish such data and link it to other datasets and re-use pre-exisiting vocabularies.
- LD also gives us access to technologies and standards such as SPARQL and SWRL (Semantic Web Rule Language) and SQRWL, an SWRL based query language.
- These standards/technologies can be used to model and then to query and work with lexico-semantic data representing meaning shift.

## LemonDia Model

- *Lemon* has become a de facto standard for publishing lexical resources as linked data
- In previous work we defined an extension of *lemon* called *lemonDia* that enables the addition of temporal information
- The main idea with is to add a temporal parameter to the sense relation linking together a Lexical Entry and a Lexical Sense.
- We chose to treat lexical senses as perdurants with an inherent temporal extent, thus defining a special subclass of LexicalSense in *lemonDia* called LexicalpSense.
- Each member of LexicalpSense has an associated time period, a member of time:ProperInterval, to which it is linked via the *lemonDia* property temporalExtent.

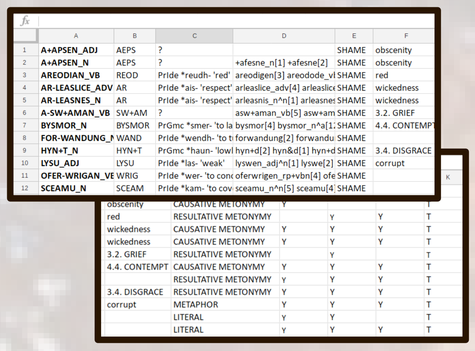



## Focus on Temporal Aspects

- The OWL-Time ontology allows us to express facts about topological relations among instants and intervals, together with information about durations, and about temporal position in RDF.
- When we don't have specific dates, we can use Allen's basic relations to define time periods in terms of their relations to each other (Allen, 1983), that is qualitatively instead of quantitatively, see (Batsakis et al., 2009).

## The Dataset

- The examples are taken from a lexical dataset of Old English (OE) emotion terms that was produced as part of a wider study into the cognitive implications of meaning change by Diaz-Vera (Diaz-Vera 2014)

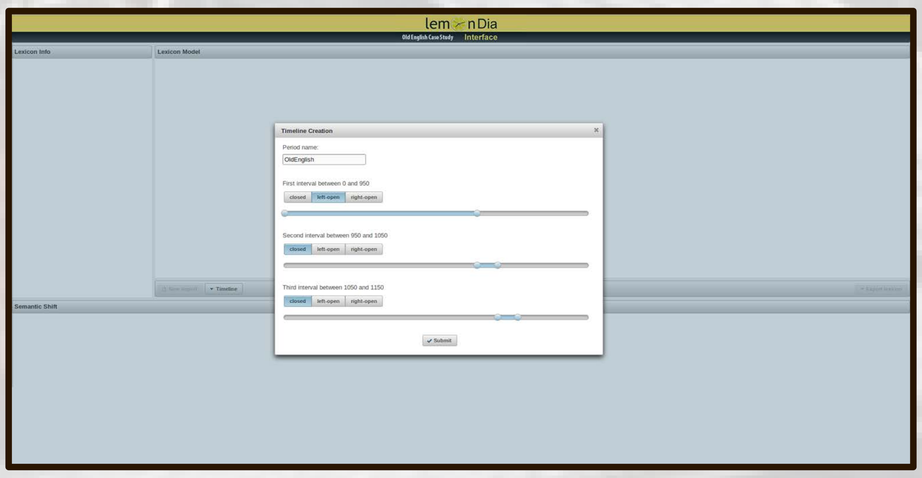


| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **A+APSEN_ADJ** | AEPS | ? | | SHAME | obscenity |
| 2 | **A+APSEN_N** | AEPS | ? | +afesne_n[1] +afesne[2] | SHAME | obscenity |
| 3 | **AREODIAN_VB** | REOD | Pride *reudh- 'red' | areodigen[3] areodode_vl | SHAME | red |
| 4 | **AR-LEASLICE_ADV** | AR | Pride *ais- 'respect' | arleaslice_adv[4] arleaslic | SHAME | wickedness |
| 5 | **AR-LEASNES_N** | AR | Pride *ais- 'respect' | arleasnis_n^n[1] arleasnes | SHAME | wickedness |
| 6 | **A-SW+AMAN_VB** | SW+AM | ? | asw+aman_vb[5] asw+am | SHAME | 3.2. GRIEF |
| 7 | **BYSMOR_N** | BYSMOR | PrGmc *smer- 'to la | bysmor[4] bysmor_n^n[1] | SHAME | 4.4. CONTEMPT |
| 8 | **FOR-WANDUNG_N** | WAND | Pride *wendh- 'to t | forwandung[2] forwandu | SHAME | |
| 9 | **HYN+T_N** | HYN+T | PrGmc *haun- 'low | hyn+d[2] hyn&d[1] hyn+d | SHAME | 3.4. DISGRACE |
| 10 | **LYSU_ADJ** | LYSU | Pride *las- 'weak' | lyswen_adj^n[1] lyswe[2] | SHAME | corrupt |
| 11 | **OFER-WRIGAN_VB** | WRIG | Pride *wer- 'to cov | oferwrigen_rp+vbn[4] ofe | SHAME | |
| 12 | **SCEAMU_N** | SCEAM | Pride *kam- 'to cov | sceamu_n^n[5] sceamu[4] | SHAME | |

| | | | | | |
|---|---|---|---|---|---|
| obscenity | CAUSATIVE METONYMY | Y | | | T |
| red | RESULTATIVE METONYMY | | Y | Y | T |
| wickedness | CAUSATIVE METONYMY | Y | Y | Y | T |
| wickedness | CAUSATIVE METONYMY | Y | Y | Y | T |
| 3.2. GRIEF | RESULTATIVE METONYMY | | Y | | T |
| 4.4. CONTEMPT | CAUSATIVE METONYMY | Y | Y | Y | T |
| | RESULTATIVE METONYMY | Y | Y | Y | T |
| 3.4. DISGRACE | RESULTATIVE METONYMY | Y | Y | Y | T |
| corrupt | METAPHOR | Y | Y | Y | T |
| | LITERAL | Y | Y | | T |
| | LITERAL | Y | Y | Y | T |

## The Interface

- The interface allows the user to specify the time periods involved in the dataset, using a slider to create time intervals.
- A lexicon is then generated using the data of the input file
- Afterwards it is possible to export this lexicon in various formats, such as RDF/XML and TURTLE, or JSON, in order to use it as input to one of the stages in a NLP pipeline.




- The interface is composed of three panels:
- The "lexicon info" panel
- the "model" area contains the lexicon in *lemonDia*
- the "semantic shift" area shows the temporal evolution

## Querying

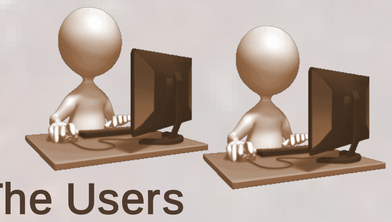
- After the interface has produced an output we can query it using the SQRWL querying language using rules which work with qualitative intervals such as those published by Batsakis. The following two queries show the kinds of queries which we can formulate on the dataset.

**Senses that have a temporal extent of OE1**

*ld:LexicalpSense(?x) ^ ld:temporalExtent(?x,anglo: OE1) -> sqwrl:select(?x)*

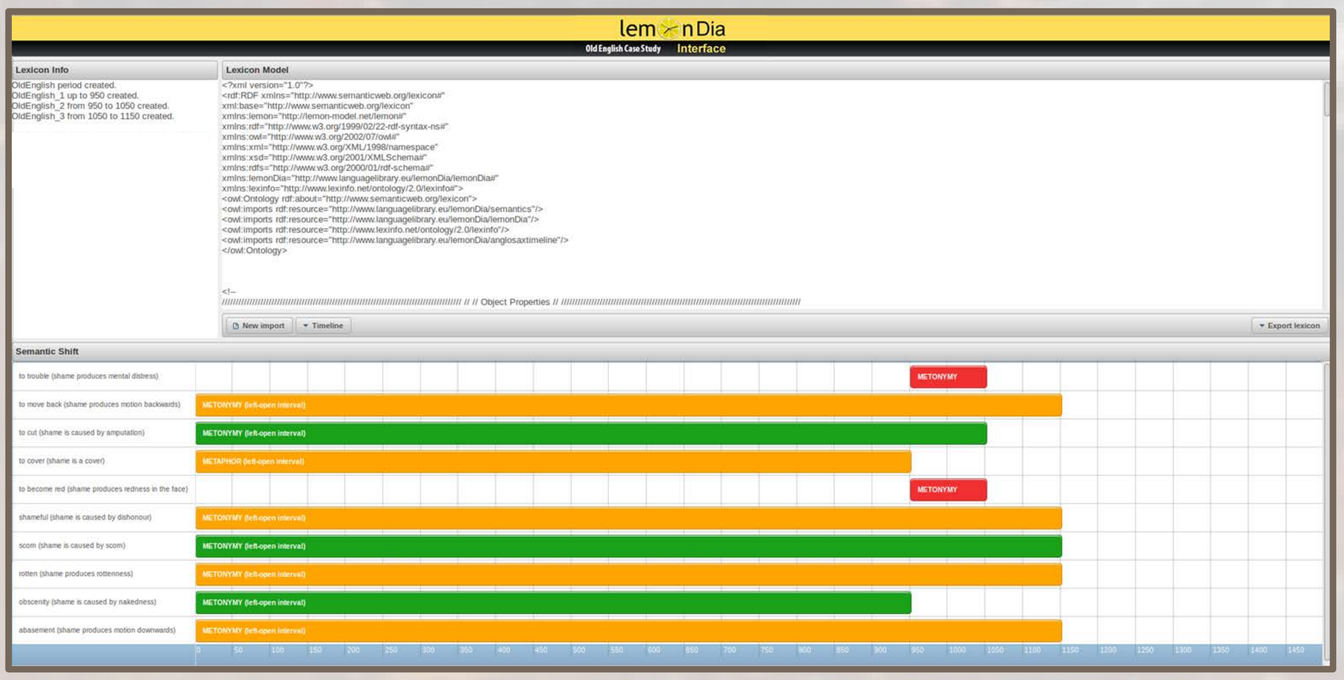
**Senses contained in OE1 time period**

*ld:LexicalpSense(?x) ^ ld:temporalExtent(?x,?y) ^ intervalContains(?y, anglo:OE2) -> sqwrl:select(?x)*



## The Users

- We developed the interface to help non-expert users, in particular lexicographers and linguists, to create RDF datasets using the *lemonDia* model since it can be quite difficult to work directly with RDF.




## Data Entry

- The interface then takes as input an Excel or CSV file containing the lexical entries in the dataset.

| column | info |
|---|---|
| a | lexical entry |
| b | root entry |
| c | etymology |
| d | lemmas |
| ... | |
| i | semantic shift target |
| j | semantic shift source |
| ... | |
| m,n | type of shift |
| ... | |
| q,r,s | time period |

## References

James F Allen. 1983. Maintaining knowledge about temporal intervals. Communications of the ACM, 26(11):832– 843
Sotiris Batsakis, Euripides Petrakisb, Ilias Tachmazidisa, and Grigoris Antonioua. 2009. Temporal representation and reasoning in owl 2.0. Semantic Web Journal.
Javier E Dıaz-Vera. 2014. From cognitive linguistics to historical sociolinguistics: The evolution of old english expressions of shame and guilt. Cognitive Linguistic Studies, 1(1):55–83.
Fahad Khan, Federico Boschetti, and Francesca Frontini. 2014. Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources. Proceedings of the Workshop on Linked Data in Linguistics 2014 (LDL-2014).
Fahad Khan, Javier E. Dıaz-Vera, and Monica Monachini. 2016. Representing polysemy and diachronic lexicosemantic data on the semantic web. In Proceedings of the 2° International Workshop on Semantic Web for Scientific Heritage
John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asuncion Gomez Perez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2010. The lemon cookbook.
Martin O'Connor and Amar Das. 2009. Sqwrl: A query language for owl. In Proceedings of the 6th International Conference on OWL: Experiences and Directions - Volume 529, OWLED'09, pages 208–215, Aachen, Germany, Germany. CEUR-WS.org.