WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation

Tomasz Naskręt¹, Agnieszka Dziob¹, Maciej Piasecki¹, Chakaveh Saedi², António Branco²

¹G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Science and Technology, Wrocław, Poland
& CLARIN-PL clarin-pl.eu

²University of Lisbon, Faculty of Sciences, Department of Informatics, Portugal
NLX-Natural Language and Speech Group
Agenda

- Context and goal: a wordnet editor
- Basic assumptions for a wordnet editor
- Graph-based presentation
- Architecture
- Extensions and Applications
  - plWordNet development
  - Portuguese Wordnet
- Conclusions and Further Works
Context and Goal

Context

- A wordnet is a complex graph of several types of nodes and edges
- WordnetLoom 1.0: simultaneous browsing and editing wordnet graphs
- Limitations: focus on monolingual wordnet and a quite inefficient thick client model

Goal

- a new re-built and expanded, version of WordnetLoom 2.0
- based on an efficient software architecture of a thin client
- facilitating work on a multilingual system of wordnets and more flexibility in enriching wordnet representation
- discussion of its applications and variants, e.g. for *MultiWordnet of Portuguese*
Basic Assumptions for a Wordnet Editor

- All editing actions should be done only via GUI
- Support for distributed group work on the central database
- Corpus-based wordnet development paradigm
  - extraction of the most frequent lemmas from a large corpus
  - corpus-based a measure of semantic similarity
  - clustering lemmas into *packages* – units of work assignments
- Substitution tests – intrinsic parts of the relation definitions to be stored and presented
- A relation graph is the basic means for both browsing and editing the wordnet structure
  - the user can freely browse the network unfolding as many levels and parts as he wants
  - *direct editing* – every link can be added or removed directly on the graph
Basic Assumptions for a Wordnet Editor

- Construction of the mappings between wordnets should be also based on visual graph presentation
  - wordnets for different languages presented simultaneously on the screen as graphs
  - inter-lingual relations visually shown on the screen
  - *direct multilingual editing*

- Non-relational elements of description
  - e.g.: glosses, usage examples, and different attributes, e.g. stylistic register, sentiment polarity
  - different perspectives: not only graph-based, but also more dictionary-oriented
  - perspectives on data: *perspective of lexical units, visualisation and synsets*
Graph-based Presentation

Assumptions

- Two types of wordnet relations
  - relations expressing some aspects of hierarchy (e.g. hypernymy/hyponymy, type-instance)
  - other relations (e.g. holo/meronymy)
- Inadequacy of typical presentation schemes, e.g.
  - radial: characteristic features of the hierarchical relations are lost
  - tree-like: the majority of its relations do not form a tree
- Unique combination of the radial and tree-like presentation
  - structure relations are presented along the vertical dimension
  - other relations are presented radially around synsets
- User initiated exploration: unfolding and browsing many levels, presentation of links on demand
Graph-based Presentation

Example
Graph-based Presentation

Example: hiding links
Graph-based Presentation
Example: expanding hidden links
Double layer graph: synsets and lexical units as nodes
- cross-linked: lexical units are synset members
- two inter-connected graphs is too much for one screen

Only the synset graph is visually presented
- synset in focus
- lexical units and their relations are presented in a separate side panel

Large synsets: less than 2 on average, but up to 20
- more important to see the structure
- only one synset member, the first lexical unit presented in the graph
- full list of lexical units in a side panel
Graph-based Presentation

Combined graphs
Graph-based Presentation

Bird eye view
Combined graphs
Example: Synset presentation

Lexical units in synset:

- **samochód 1 (wytw)**
- auto 1 (wytw)
- wóz 1 (wytw)
- pojazd samochodowy 1 (wytw)

Comment:

n/d

Synset Id: 7337
Owner: n/d
Combined graphs
Example: lexical unit properties
Combined graphs
Example: lexical relations
Experimental Graph of Lexical Relations
Architecture
Scheme of the platform

Presentation Tier
- WordnetLoom (Java Swing Thin Client)
- Słowosieć Web Application
- Statistic (Web front)

Business logic Tier
- Validation
- CRUD operations
- Statistics
- Monitoring

Data Tier
- Database
Architecture
Selected features

- Java-based implementation
  - free of the problems related to the changing versions of web-browsers
  - works on every operating system
  - easy to install by non-technological users
- Based on *MySql 5.7* database management system
- *Hibernate Envars* module allows for easier undoing of changes
- Database schema has been rebuilt to be similar to the UBY-LMF structure
- All dictionaries are stored in the database; it supports localisation mechanisms
- Users can choose which lexicons, mostly wordnets, they want to work with
- Extensible validation module to prevent errors including some semantic errors
Rich experience collected during more than 10 years of using WordnetLoom for plWordNet editing (> 50 person-years)

Multilinguality
- inter-lingual relations are synset relations, but between synsets in different languages
- any number of wordnets for any number of languages can be edited on the same screen

Additional status meta-attribute and support for team management
- editors are assigned packages of lemmas and are obliged to identify and add all lexical units
- *not processed* (default value), *error*, *verified*, *new*, *partially processed*
- *added sense* – a lexical unit added from the outside of a package
Improved navigation
- search function was also expanded to cover all attributes
- navigation: a synset $\leftrightarrow$ a lexical unit

Improved diagnostics
- PoS tags to variables in substitution tests $\rightarrow$ automated control of the link correctness
- easier adding new types of lexicographic files and annotation with semantic domains
Extensions and Applications
Using WordnetLoom in Portuguese MultiWordNet (1/2)

Enhancement in

- Wordnet content
  - Language variants
    1. specific spellings (e.g. receção and recepção)
    2. specific words (e.g. autocarro and ônibus)
    3. specific syns (e.g. camisola: t-shirt or nightdress)
  - Mapping to SUMO ontology

- Lexicographer work
  1. new labels for senses/synsets (e.g. ”unchecked”, ”checked”)
  2. more search options, including by the new labels
Extensions and Applications
Using WordnetLoom in Portuguese MultiWordNet (2/2)

Enhancement in
- Format compatibility
  - converter WNPrincet (syns-based) to WNLoom (sense-based)
  - any Princeton-convertible WN is now loadable into WNLoom
- Technical issues
  - bugs with words with multiple senses
  - bugs in the GUI
  - other issues
Conclusions and Further Works

- **WordnetLoom** incorporates more than 10 years of experience in the development of a very large wordnet by many linguists on a daily basis.
- This rich experience has become a good basis for the development of new version improved with respect to both: technology and functionality.
- **WordnetLoom** is open: https://github.com/CLARIN-PL/WordnetLoom
- Most unique features
  - direct work on the visually presented wordnet graph
  - simultaneous editing and inter-linking of many wordnets
- Adaptation for Portuguese Wordnet developed according to completely different method
- Further collaborative development of the system
Thank you very much for your attention!

http://clarin-pl.eu
http://nlp.pwr.edu.pl
http://plwordnet.pwr.edu.pl
https://github.com/nlx-group