Lexicographic Perspective on Wordnet Interoperability in CLARIN

Maciej Piasecki, Ewa Rudnicka, Tomasz Walkowiak
G4.19 Research Group, Wrocław University of Science and Technology

Darja Fišer
Faculty of Arts, Univerza v Ljubljani

wnlex Workshop, Ljubljana 2018-07-16
Plan

• CLARIN in a snapshot
• Diversified lexical resources in CLARIN
• Applications vs interoperability
• Idea of federated search across lexical resources,
• Lexical Platform – a light way solution for users
• Perspectives on deeper, technological integration
CLARIN ERIC - *Common Language Resources and Technology Infrastructure*

- CLARIN is ERIC type consortium of 20 members and 2 observers: countries and international organisations
- Focus area: supporting research in Humanities and Social Sciences
- CLARIN Mission
  - to collect language resources and tools for languages used in Europe in one shared, distributed infrastructure with uniform access
  - to significantly lower the barriers for the use of Language Technology in Humanities & Social Sciences (H&SS)
  - to facilitate or enable research methods based on automated analysis of text and speech resources
CLARIN ERIC (European Research Infrastructure Consortium)

- 20 members:
  - Austria
  - Bulgaria
  - Croatia
  - Czech Republic
  - Denmark
  - Dutch Language Union
  - Estonia
  - Finland
  - Germany
  - Greece
  - Hungary
  - Italy
  - Latvia
  - Lithuania
  - The Netherlands
  - Norway
  - Poland
  - Portugal
  - Slovenia
  - Sweden

- Observers:
  - France, Great Britain
Language Technology for H&SS

• Language Technology (LT)
  - language resources and tools
  - robust in terms of quality and coverage,
  - multipurpose
  - component based

• Language Technology Infrastructure
  - a software framework (architecture or platform)
  - for combining language tools with language resources into processing chains (or pipelines)
  - the defined processing chains are next applied to language data sources
  - interoperability, also with the external systems
Language Technology for H&SS

• Limited usage of LT in Humanities and Social Sciences
  - hard to find: dispersed in the Web, poorly described in a technical language
  - varieties of technological solutions, insufficient users’ computers
  - required programming skills or knowledge from the area of natural language engineering

• LT Infrastructure for H&SS
  - common standards, combined platforms, open approaches
  - joint catalogues and search facilities
  - focused on H&SS users and support for them
  - Web Services and Web Applications: no need for installing, processing focused on H&SS research tasks
Distributed Infrastructure
CLARIN is a persistent, distributed system
CLARIN: Central Services

https://www.clarin.eu/

**CLARIN portal**
Get an example-based impression of what's currently available

**Depositing services**
Store language resources in a sustainable repository at a CLARIN centre

**Virtual Language Observatory**
Discover language resources using a faceted browser or a map

**Easy access to protected resources**
Get easy access to protected resources, with your institutional username and password.

**Web services and applications**
Explore and analyze language data with a wide variety of tools

**Virtual Collections**
Create your own digital bookmarks, ideal for citing data sets.

**Language Resource Inventory**
Submit and access information about language resources relevant to your

**Content Search (prototype)**
Search different corpora with a single search engine

**Questions**
Searching for a specific data set or application? Wondering how CLARIN can
CLARIN Basic functions

• Facilitating access to language resources
  - federation of repositories - Virtual Language Observatory
  - federated search across corpora - Federated Content Search

• Support for automated analysis of text and speech
  - a range of ready to use language tools
    • Web Services and (web) applications
    • Access through repositories

• Research applications
  - built for concrete needs, often in cooperation with users (researchers)
  - Based on LT, but not ‘imposing’ it on users
Virtual Language Observatory

VLO / Faceted search / Search results

Showing 1 to 10 of 88 results for wordnet

Use the categories below to limit the search results to those matching the selected value(s).

Language
Collection
Resource type
Modality

Type to search for more

writtenlanguage (5)
written (4)
multimodal (1)
other (1)

Format

WordNet

plWordNet is a lexico-semantic network which reflects the lexical system of the Polish language. There are at present ca. 144,000 nouns, verbs and adjectives in plWordNet, ca. 203,000 word senses and ca. 500,000 relations. It is already the second-largest wordnet in the world, and it keeps growing.

Finnish WordNet

The language resource is available in Kielipankki - the Language Bank of Finland at http://urn.fi/urn:nbn:fi:lb-2016042205; download: http://urn.fi/urn:nbn:fi:lb-2014052713. The Finnish WordNet is a lexical database for Finnish. It is a part of the FIN-CLARIN infrastructure project. FinnWordNet is lic...

Czech WordNet

Multilingual Lexicons; The Czech WordNet was developed by the Centre of Natural Language Processing at the Faculty of Informatics, Masaryk University, Czech Republic. The Czech WordNet captures nouns, verbs, adjectives, and partly adverbs, and contains 28,201 word senses (synsets). Every synset encodes the equivalenc...
VLO – Facet-based Search and Browsing

• Common meta-data standard:
  • CMDI
  • = Component Metadata Infrastructure
• Facet-based search browsing
  • Fields and values acquired from CMDI records
  • Semantically grouped
CMDI Components

• profile

• name
• contact
• ...

• name surname
• nationality
• ...

• name
• id
• ...

• format
• size
• ...

project

author

language

technical metadata
Diversified lexical resources in CLARIN

• Types:
  - word lists and frequency lists
  - gazetteers (lists of Proper Names)
  - morphological dictionaries
  - bilingual word-to-word dictionaries
  - dictionaries with textual descriptions
  - grammatical lexicons:
    • valency frames, some with semantic restrictions

• Wordnets:
  - based on Princeton WordNet – transfer method
  - manually constructed from scratch
  - wordnet-like semantic lexicons
Wordnets in CLARIN ERIC infrastructure

• Wordnets, e.g.:
  - based on Princeton WordNet by transfer
    • Princeton WordNet, Ancient Greek WordNet, Estonian WordNet, Finnish WordNet, ItalWordNet (Italian), MultiWordNet (Portuguese, Italian), sloWNet (Slovene), ...
  - manually constructed from scratch
    • DanNet (Danish), GermaNet (German), plWordNet (Polish), enWordNet (English)
  - wordnet-like semantic lexicon
    • Saldo (Swedish), Swesaurus (Swedish)
Applications of wordnets

• H&SS users’ manual work
  - dictionary consulting: single entry, examples, comparing and contrasting resources,
  - thematic lists of entries
  - classification of words and meanings
  - language comparison

• Language processing
  - Word Sense Disambiguation
  - expanded text representation
  - semantic frequency text representation
  - thematic text classification and grouping
Interoperability - problems

• Models – understanding of basic notions
  - synsets, lexical units,
  - relation types and subtypes,
  - glosses, usage examples, comments,
  - frames, attributes, restrictions,
  - stylistic registers, domains (representation & definition), contexts,
  - emotive description (e.g. emotions, sentiment polarity) etc.

• Models - correspondence:
  - between basic building blocks: anchor points
  - direct and indirect mapping between wordnets
  - linking with other resources, e.g. LOD
Interoperability - problems

• Wordnets and knowledge sources (knowledge databases)
  - especially Linked Open Data and terminological databases
  - strategies for mapping, anchor points
  - division of material

• Multilingual mapping
  - types, models, languages
    • synsets vs lexical units (senses)
  - cross-lingual anchor points
plWordNet on WordNet mapping procedure

- Recognise the sense of a source synset by:
  - its position in the network structure,
  - existing relations, commentaries (glosses),
  - comparison to other synsets containing the given lemma

- Search for candidates for a target synset:
  - intuitions, automatic prompting and dictionaries

- Verify candidates:
  - by comparing hypernymy and hyponymy structures
  - by exploring existing inter-lingual relations;
  - by comparing definitions, commentaries; dictionaries

- Link the source synset with the target synset
plWordNet mapping in WordnetLoom system
plWordNet mapping: problems

- Lexico-grammatical differences
  1. Markedness
     - young being (prosiak 'piglet' - hypo → młodzik 'young creature')
     - diminutive (prosiaczek 'piggy' ← prosiak + -ek)
     - augmentative
  2. Lexicalised gender
     - e.g. cousin ~ kuzyn (masc.) & kuzynka (fem.)
  3. Lexical gaps
     - E.g. names of different types of relatives
     - {kaowiec 1}, a Polish term denoting an institution’s employee responsible for the organization of cultural and recreational events in the Communist times
plWordNet mapping: problems

• Differences in synonymy and the structure of synset:
  4. Mixed Princeton WordNet synsets:
     a) neutral and marked forms in the same synset
     b) gender, e.g. {bondswoman 1, bondsman 1}
     c) countability - mass/count distinction, | e.g. {furniture 1, piece of furniture 1}
     d) hypernym and hyponyms in one synset e.g. {monte 1, three card monte 1}
     e) multiple hypernymy and vs. and/or

• Other differences:
  5. Glosses contrary to relations
  6. Different relations to code the same conceptual dependencies: meronymy vs. hyponymy
  7. More detailed sense distinctions (&) Dictionary content mismatches...
plWordNet mapping: inter-lingual relations

- In total: >274,000 links from plWordNet to WordNet 3.1
- Cascade of noun relations – the first which fits is chosen

<table>
<thead>
<tr>
<th></th>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Inter-lingual synonymy</td>
<td>38,674</td>
</tr>
<tr>
<td>2.</td>
<td>Inter-lingual inter-register synonymy</td>
<td>1,848</td>
</tr>
<tr>
<td>3.</td>
<td>Inter-lingual partial synonymy</td>
<td>5,747</td>
</tr>
<tr>
<td>4.</td>
<td>Inter-lingual instance</td>
<td>616</td>
</tr>
<tr>
<td>5.</td>
<td>Inter-lingual partial hyponymy</td>
<td>82,033</td>
</tr>
<tr>
<td>6.</td>
<td>Inter-lingual hypernymy</td>
<td>29,336</td>
</tr>
<tr>
<td>7.</td>
<td>Inter-lingual meronymy</td>
<td>10,756</td>
</tr>
<tr>
<td>8.</td>
<td>Inter-lingual holonymy</td>
<td>7,816</td>
</tr>
<tr>
<td>9.</td>
<td>Inter-lingual type</td>
<td>7,505</td>
</tr>
</tbody>
</table>
plWordNet mapping: inter-lingual relations

- Cascade of adjective and adverb relations – the first which fits is chosen

1. Inter-lingual synonymy 4,491 998
2. Inter-lingual inter-register synonymy 97 48
3. Inter-lingual partial synonymy 1,600 311
4. Inter-lingual cross-categorial synonymy 24,715 98
5. Inter-lingual partial hyponymy 44,206 9,842
6. Inter-lingual hypernymy 311 112
Interoperability - problems

• Accessibility and compatibility of licenses
  - different licences
    • open – restricted academic – commercial
  - problems in combining resources

• Aggregators and credits for the authors
  - aggregated resources: different languages, different types
    • e.g. BabelNet, Open Multilingual WordNet
  - limited recognition of the aggregated individual wordnets by users
  - citations directed to aggregates instead to the creators of the resources
Interoperability - formats

• Common format for lexico-semantic resources – almost a mission impossible
  - native XML formats, LMF, TEI, RDF, ...
• Wordnet formats
  - native XML-based, e.g. DEBVisDic
  - RDF-based, e.g. WordNet RDF in DanNet
  - LMF, e.g. KyotoLMF, GermaNet LMF, UBY LMF but also **CILI LMF**
  - Lemon (with conversion to CILI LMF)
• Versions, granularity and Persistent Identifiers
Interoperability - attempts

• Multilingual Central Repository and WordTies
  - common format, common database, loss of information

• BabelNet
  - common format, common database, loss of information
  - aggregator with very limited visibility of components, very restricted licence

• Open Multilingual WordNet and CILI
  - common format focused only on wordnets
  - aggregator, limited to open resources
Idea of federated search across lexical resources

• Idea
  - a virtual place for aggregating different types of LRs as separate individual components
  - in a way that they form an interconnected system, a complex LR, from the user point of view

• Fundamental assumptions
  - descriptions provided for LRs must be minimal
  - no common format required for the full usability of the platform by non-technological users
  - solution open for all types of LRs
    • with a special focus for wordnets
  - all LRs must preserve their identity
Federated search across LRs: basic assumptions

- Independence of LRs
  - different LRs are grouped as independent components, implemented as software modules
  - minimal requirements are imposed on developers
  - individual identity of all LRs must be visible and preserved
  - distribution
    - a component may be located in any freely selected network location
    - no need to copy LR data to the platform – IPR protection

- Format agnostic
  - a limited set of common formats promoted
  - any specific data format not imposed on the platform on its components
Federated search across LRs: basic assumptions

• Data encapsulation
  - a component will be only accessible via a set of Programming Interfaces (PIs), e.g. Web Services
    • no means for a direct access to LR data
  - one PI can be implemented as a one separate WS, or several PIs can be provided by a single WS
  - only a minimal set of PIs is required to be implemented by every component
  - all components can provide any number of additional PIs

• Read-only access
  - Lexical Platform is a tool for accessing a complex system of linked resources
  - no editing, updating will not be supported
Federated search across LRs: basic assumptions

• Inter-linking
  - based exclusively on the content of LRs
  - each component recognises references to elements of the limited set of types:
    • selected points by which the data from different components are anchored to the whole platform and inter-linked between them
  - anchor elements should naturally originate from the construction of a LR
    • characteristic elements for browsing
    • native (or natural, typical) mapping to other LR
  - expected types for inter-linking
    • word form (including multi-word expressions), lemma (literals, entry form,...), lexical unit (word sense), synset (id), frame (syntactic and/or semantic), domain (context), and concept
Federated search across LRs: functions

• Learning about the component LRs and the range of information provided by them, e.g. by a query;
• Searching across combined LRs on the basis of anchor elements supported by different components
• Browsing LRs by lists of anchor elements retrieved from the components
• Navigating across linked LRs on the basis of anchor elements;
• Finding out how to access and download original resources and obtain information on how to browse different LRs in their native browsers
Lexical Platform a light way solution for users

User query: „to postpone”

HTML
Synsets, lemmas, relation links, (anchor points)

HTML
senses, syntactic frames, restrictions, (anchor points)

WordnetA
Present

VerbnetA
Present

Registered components
WordnetB
Framenet
Thesaurus
Wiktionary
…

Anchor point
Lexical Platform a light way solution for users
http://lexp.clarin-pl.eu
Lexical Platform a light way solution for users
http://lexp.clarin-pl.eu

![Open Multilingual Wordnet](image)

Other languages 🌐

<table>
<thead>
<tr>
<th>Meaning</th>
<th>List of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>a dwelling that serves as living quarters for one or more families</td>
<td>en: house</td>
</tr>
<tr>
<td></td>
<td>es: casa</td>
</tr>
<tr>
<td>aristocratic family line</td>
<td>en: house</td>
</tr>
<tr>
<td></td>
<td>es: casa</td>
</tr>
<tr>
<td>where you live at a particular time</td>
<td>en: home, place</td>
</tr>
<tr>
<td></td>
<td>es: casa</td>
</tr>
</tbody>
</table>
Lexical Platform a light way solution for users
http://lexp.clarin-pl.eu

DOM (7018) sb m
dóm (151), dom (84), dom- (182), dóm- (15); dóm: ForCnR, BialKat (6), Strum, KochMarz, KochPieś (4), KochWr (2), SiebRozmyśli (2), GasCast (2); dom: KierPow, KochTr, KochFr (5), KochDe (2), Ostrégii (2), JanNKar, JanNKarKoch; dóm: dom Opeczyń (4 : 28), MurzNT (1 : 3), Męcz (139 : 1), OrzQuin (1 : 3), KochPs (1 : 10), KochWz (1 : 2), KochMRot (1 : 1), ZawKef (2 : 2), KochFrag (2 : 5), SarnStat (20 : 14); dóm: Strum; dom- : dóm-, BialKat (9 : 6), KochPieś (12 : 1), PudlFr (8 : 2), GórniTros (14 : 3), KochFrag (11 : 1), SarnStat (130 : 1).

<table>
<thead>
<tr>
<th>Fleksja</th>
<th>sg</th>
<th>pl</th>
<th>du</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>dóm</td>
<td>domy</td>
<td>domu</td>
</tr>
<tr>
<td>G</td>
<td>domu, dom</td>
<td>domów</td>
<td>domu</td>
</tr>
<tr>
<td>D</td>
<td>domowi, domu</td>
<td>domów</td>
<td>domu</td>
</tr>
<tr>
<td>A</td>
<td>dóm</td>
<td>domy</td>
<td>domu</td>
</tr>
<tr>
<td>I</td>
<td>domem</td>
<td>domami</td>
<td>domami</td>
</tr>
<tr>
<td>L</td>
<td>domu, domem</td>
<td>domach, domach</td>
<td>domach, domoch</td>
</tr>
<tr>
<td>V</td>
<td>domiech, domem</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sl sfp, Cn notuje, Linde XVI – XVIII w.

Znaczenia
1. Budynek, najczęściej budynek mieszkalny; mieszkanie, stałe miejsce zamieszkania
   Przen
   a) Niebo, raj jako mieszkanie Boga, przyszłe mieszkanie wieczne człowieka
   b) Serce, dusza, ciało człowieka jako mieszkanie Boga lub Ducha świętego; ciało jako mieszkanie duszy
2. Budynek przeznaczony na jakiś cel; zakład, instytucja
   a. Świątynia, miejsce modlitwy
   Przen: Kościół jako instytucja, ogół wiernych
3. Rodzina, mieszkańcy domu, ognisko domowe; gospodarstwo; majątność, posiadłość
4. Ród, dynastia; pochodzenie
5. Strony rodzinne: król, ojczyna, siedziba
Lexical Platform a light way solution for users
http://lexp.clarin-pl.eu

Znaczenia
1. Budynek, najczęściej budynek mieszkalny; mieszkanie, stałe miejsce zamieszkania
   Przen
   a) Niebo, raj jako mieszkanie Boga, przyszłe mieszkanie wieczne człowieka
   b) Serce, dusza, ciało człowieka jako mieszkanie Boga lub Ducha świętego; ciało jako mieszkanie duszy
2. Budynek przeznaczony na jakiś cel; zakład, instytucja
   a. Świątynia, miejsce modlitwy
      Przen: Kościół jako instytucja, ogół wiernych
3. Rodzina, mieszkańcy domu, ognisko domowe; gospodarstwo; małżonkowość
4. Ród, dynastia; pochodzenie
5. Strony rodzinne; kraj, ojczyzna, siedziba
6. Klasztor, dom zakonny
7. W dawnej astrologii pewna część nieba w stosunku do odpowiedniej części Ziemi
8. n-loc

*** Bez wystarczającego kontekstu

5. ojczyzna.

C/DOMA, [DOMOGOSPODARNY], DOMORODAK, DOMORODNY, DOMOSTWO

kontakt - telefon: 566 210 331, e-mail:
Liczba odwiedzin: 517085
Lexical Platform a light way solution for users
http://lexp.clarin-pl.eu

~ Similarity ~

ogrod  dom_rodzinny  chatka  oberza  pokoj  willa  gospoda  chałupa  domostwo
palac  szopa  barak  hacienda  chat  miasto  hotel  gołębnik  szałas  stodoła  wioska
mieszkanie  karczma  izdebka  kurnik
Conclusions

• Wordnets together with semantic lexical resources form within CLARIN rich lexicographic knowledge base

• This richness is not yet enough well visible and explorable by the users

• Light way models the integration of resources can pave a road towards an interconnected network of lexical resources
Thank you very much for your attention!