CLARIN-PL
Common Language Resources and Technology Infrastructure

CENTRUM TECHNOLOGII
JĘZYKOWYCH CLARIN-PL

Annotation tasks and solutions in CLARIN-PL

Marcin Oleksy, Ewa Rudnicka
Wrocław University of Technology

marcin.oleksy@pwr.edu.pl
ewa.rudnicka@pwr.edu.pl

CLARIN ERIC

- Common Language Resources and Technology Infrastructure, European Research Infrastructure Consortium
- Resources:
  - digital archives, corpora, electronic dictionaries, and language models
- Tools for:
  - syntactic and semantic analyses, speech recognition, search for proper names or recognition of situation descriptions
- Mission:
  - interoperability of tools and resources (also from external systems)
  - resource storage, meta-data description and sharing
  - research tools for the enhanced access to large collections of source texts, spoken language records and multimedia resources, and for their automated analysis
  - a software framework (architecture or platform) for:
    - combining language tools with language resources into processing chains (or pipelines)
    - the defined processing chains next applied to language data sources

CLARIN-PL

- User-driven Language Technology Infrastructure - bi-directional approach
  - linking of Language Resources and Tools combined with the development of research applications for Humanities & Social Sciences
- Partners:
  - Wrocław University of Technology, G4.19 Research Group
  - Institute of Computer Science, Polish Academy of Science
  - Polish-Japanese Institute of Information Technology, Chair of Multimedia
  - University of Łódź, PELCRA group at Chair of English Language and Applied Linguistics
  - Institute of Slavic Studies, Polish Academy of Science
  - Wrocław University
- Main goals:
  - completing the construction of selected resources
  - building bilingual resources and specialised corpora facilitating the envisaged needs of H&SS
  - bilingual resources crucial for interoperability (priority given to Polish-English resources)
- Visit http://clarin-pl.eu/en/services/

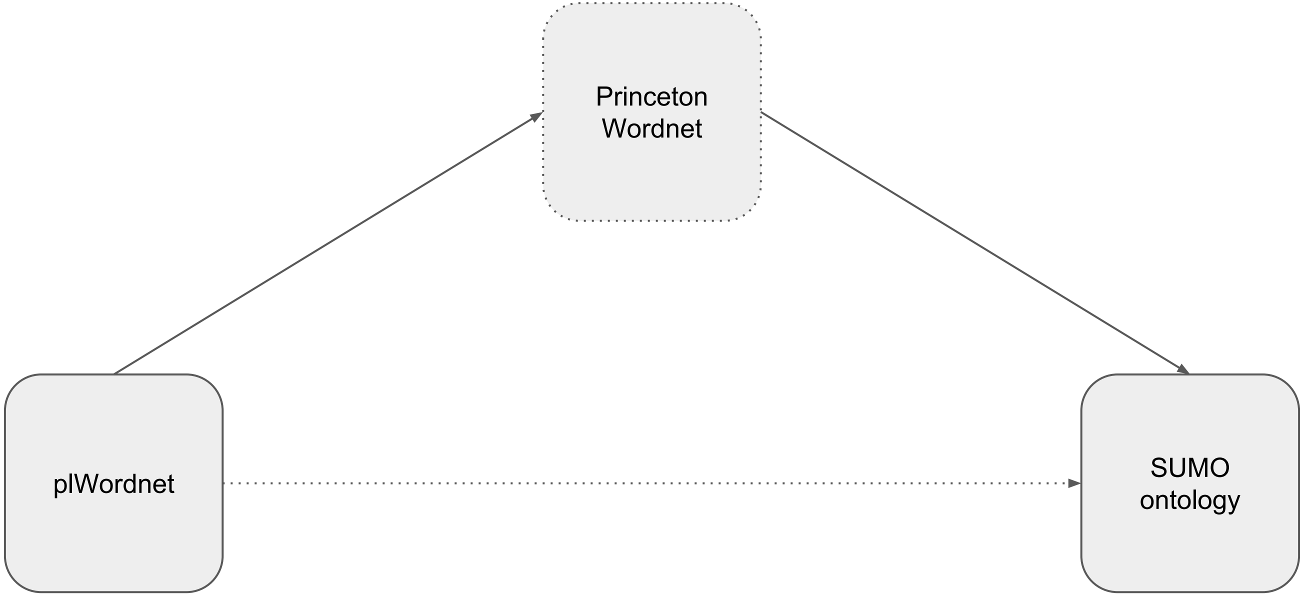
Linking monolingual resources

- Mapping plWordNet onto Princeton WordNet:
  - manual mapping supported by automatic prompt systems
  - emphasis on correspondence of wordnet structures
  - on the level of synsets (sets of synonymous lexical units (lemma - sense pairs))
- Mapping procedure:
  - reference to a variety of external sources
  - substitution tests
  - one, most informative link
- Inter-lingual relations:
  - synonymy
  - partial synonymy
  - cross-categorial synonymy
  - hypo/hypernymy
  - mero/holonymy

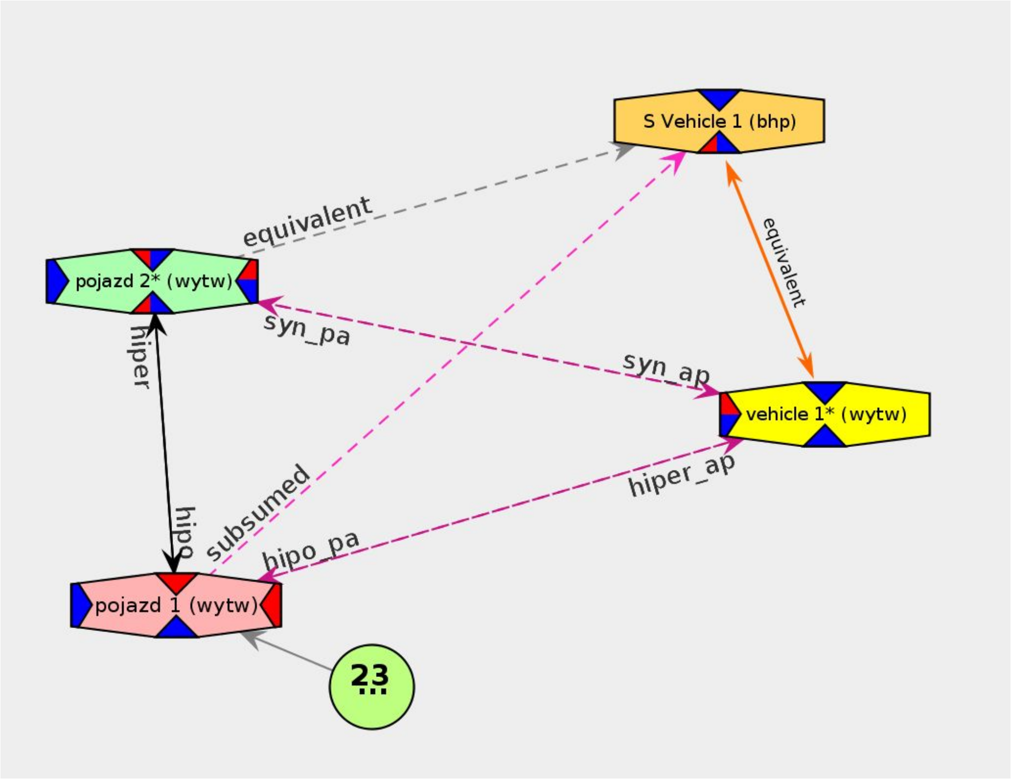
Mapping plWordnet onto SUMO ontology

- Strategy:
    - rule-based approach (about 90 rules)
    - capitalising on the existing relations:
        - plWordNet to Princeton WordNet mapping
        - Princeton WordNet to SUMO ontology mapping
        - SUMO structure
- Relations (inherited from Princeton WordNet to SUMO ontology mapping):
    - equivalent
    - instance of
    - subsumed
    - underspecified
- Results:
    - 119 000 synsets mapped onto SUMO ontology

Inter-lingual mapping as an intermediary for ontology mapping



Example of inter-lingual and ontology mapping



Annotation tasks

Annotations in KPWr 1.2

- chunks and selected predicate-argument relations
- named-entities and relations between them
- anaphora relations
- word senses
- semantic roles

Annotation tasks

Current annotation tasks:

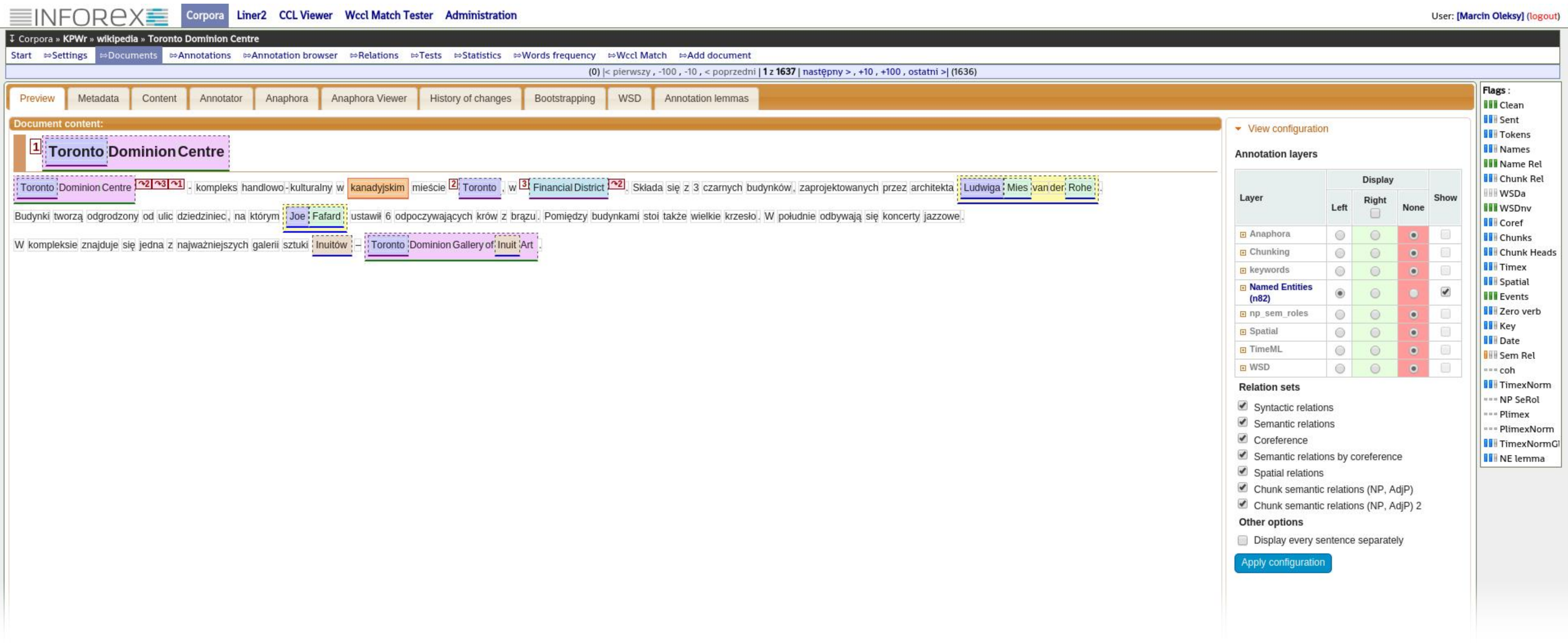
Inforex

- keywords
- temporal expressions (based on TimeML)
- events (based on TimeML)
- spatial relations (based on Spatial Role Labeling)

Sematon
(not distributed yet)

- relations between sentences (based on CST)

Annotation tasks - Inforex



User: [Marcin Oleksy] (logout)

↧ Corpora » KPWr » wikipedia » Toronto Dominion Centre

(0) |< pierwszy , -100 , -10 , < poprzedni | **1 z 1637** | następny > , +10 , +100 , ostatni >| (1636)

Document content:

1 **Toronto Dominion Centre**

Toronto Dominion Centre ↷2 ↷3 ↷1 - kompleks handlowo-kulturalny w kanadyjskim mieście 2 Toronto , w 3 Financial District ↷2 . Składa się z 3 czarnych budynków , zaprojektowanych przez architekta Ludwiga Mies van der Rohe .

Budynki tworzą odgrodzony od ulic dziedziniec , na którym Joe Fafard ustawił 6 odpoczywających krów z brązu . Pomiędzy budynkami stoi także wielkie krzesło . W południe odbywają się koncerty jazzowe .

W kompleksie znajduje się jedna z najważniejszych galerii sztuki Inuitów – Toronto Dominion Gallery of Inuit Art .

▾ View configuration

Annotation layers

Layer	Display			Show
	Left	Right ☐	None	
⊞ Anaphora	○	○	●	☐
⊞ Chunking	○	○	●	☐
⊞ keywords	○	○	●	☐
⊞ Named Entities (n82)	●	○	○	☑
⊞ np_sem_roles	○	○	●	☐
⊞ Spatial	○	○	●	☐
⊞ TimeML	○	○	●	☐
⊞ WSD	○	○	●	☐

Relation sets

☑ Syntactic relations
☑ Semantic relations
☑ Coreference
☑ Semantic relations by coreference
☑ Spatial relations
☑ Chunk semantic relations (NP, AdjP)
☑ Chunk semantic relations (NP, AdjP) 2

Other options

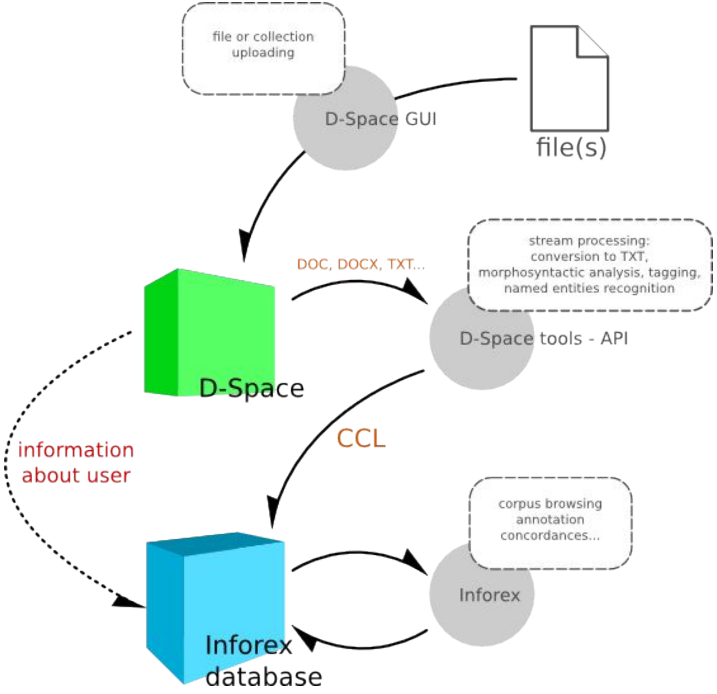
☐ Display every sentence separately

Apply configuration

Flags :

- Clean
- Sent
- Tokens
- Names
- Name Rel
- Chunk Rel
- WSDa
- WSDnv
- Coref
- Chunks
- Chunk Heads
- Timex
- Spatial
- Events
- Zero verb
- Key
- Date
- Sem Rel
- coh
- TimexNorm
- NP SeRol
- Plimex
- PlimexNorm
- TimexNormGl
- NE lemma

Annotation tasks - import process


file or collection uploading
D-Space GUI
file(s)
DOC, DOCX, TXT...
stream processing: conversion to TXT, morphosyntactic analysis, tagging, named entities recognition
D-Space tools - API
D-Space
CCL
information about user
corpus browsing annotation concordances...
Inforex
Inforex database

Annotation tasks - format

CCL format - simple format derived from XCES that allows to store:

- division into paragraphs, sentences, tokens and no-space information
- morphosyntactic and/or semantic annotations
- chunk-style annotations with possible discontinuities
- syntactic heads of annotations, properties of tokens and, implicitly, properties of annotations
- annotation channels

Various tools and resources in tools development

(spatial expressions recognition)

Goal: automatic labelling of words or phrases in sentences with a set of spatial roles which take part in one or more spatial relations expressed by the sentence.

What do we need:

- morphosyntactic patterns to identify the candidates for spatial expressions
- set of ontology-based constraints to filter out the non-spatial expressions

Various tools and resources in tools development

(spatial expressions recognition)

Information about the type of a spatial relation comes from:

- the meaning of a preposition
- meaning of lexemes referring to a localized object (trajector) and to an object of reference (landmark)

The semantic restrictions of trajector and landmark can be used to distinguish a specific meaning of the preposition due to a specific spatial cognitive schema.

Various tools and resources in tools development

(spatial expressions recognition)

Example cognitive schema

Preposition	na (on) #1
Interpretation	Object TR is outside the LM, typically in contact with external limit of LM by applying pressure with its weight.
Example	“książka leży na stole” (a book is on the table)
SUMO Class of trajector	Artifact, ContentBearingObject, Device, Animal, Plant, Pottery, Meat, PreparedFood, Chain
SUMO Class of landmark	Artifact, LandTransitway, BoardOrBlock, Boatdeck, Shipdeck, StationaryArtifact

Various tools and resources in tools development

(spatial expressions recognition)

Examples:

- TR:[{Galeria} Piastowska] w LM:[{Legnicy}]

  ‘Galeria Piastowska in Legnica’

- TR:[{trawnik}] w LM:[{parku}]

  ‘the lawn in the park’

Various tools and resources in tools development

(spatial expressions recognition)

- Galeria Piastowska w Legnicy

tagging, parsing, morphosyntactic disambiguation, chunking, named entities recognition ↓

- [{Galeria} Piastowska] [w {Legnicy}]
  [{Galeria} Piastowska] = nam_fac_goe
  {Legnica} = nam_loc_gpe_city

syntactic candidates detection ↓

- [{Galeria} Piastowska] [w {Legnicy}] = <FirstNG|...|PrepNG> (P20 syntactic pattern)

SUMO classes identification ↓

- [{Galeria} Piastowska] = Group, Transitway, StationaryArtifact, Balcony, Region, Collection, ShoppingMall, Room, RetailStore, building
  {Legnica} = City

cognitive schema matching ↓

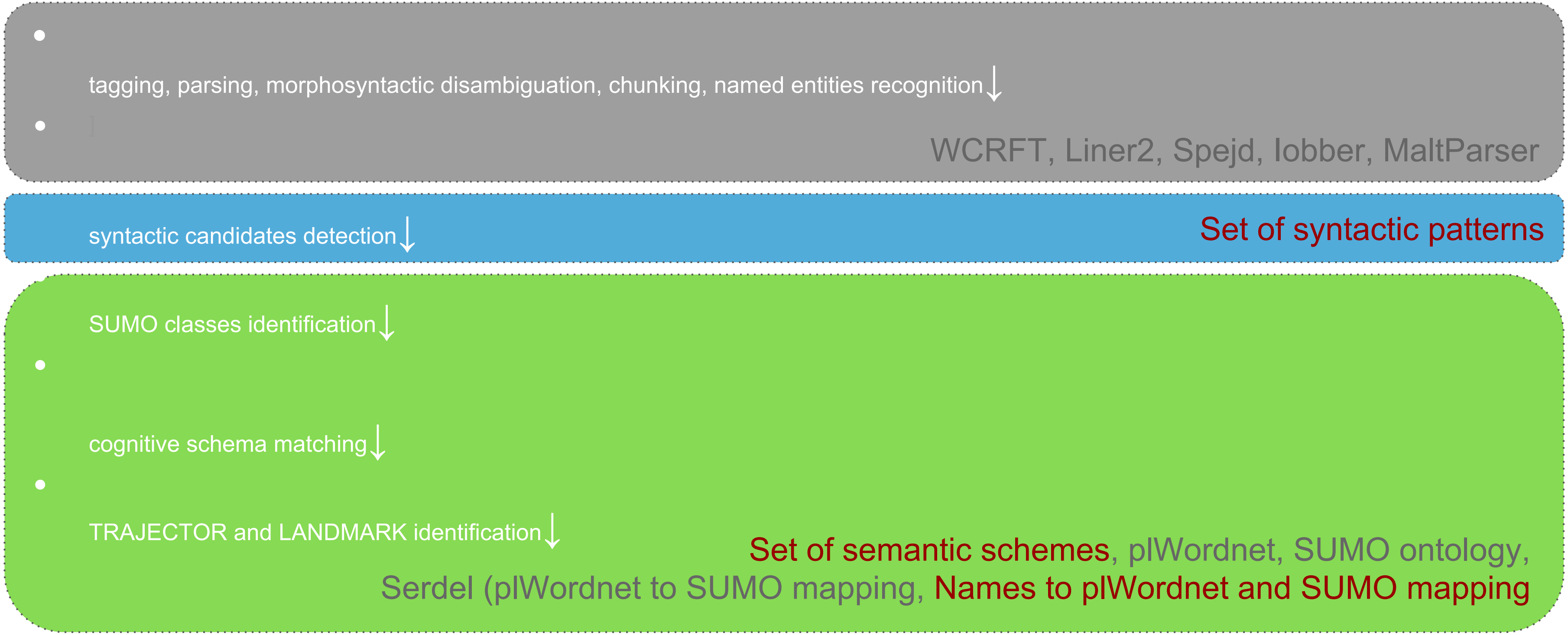
- [Group, Transitway, StationaryArtifact, Balcony, Region, Collection, ShoppingMall, Room, RetailStore, building] w [City] = w,we#w1 schema

TRAJECTOR and LANDMARK identification ↓

- TR:[{Galeria} Piastowska] w LM:[{Legnicy}]

Various tools and resources in tools development

(spatial expressions recognition)



- tagging, parsing, morphosyntactic disambiguation, chunking, named entities recognition ↓
- WCRFT, Liner2, Spejd, Iobber, MaltParser

syntactic candidates detection ↓

Set of syntactic patterns

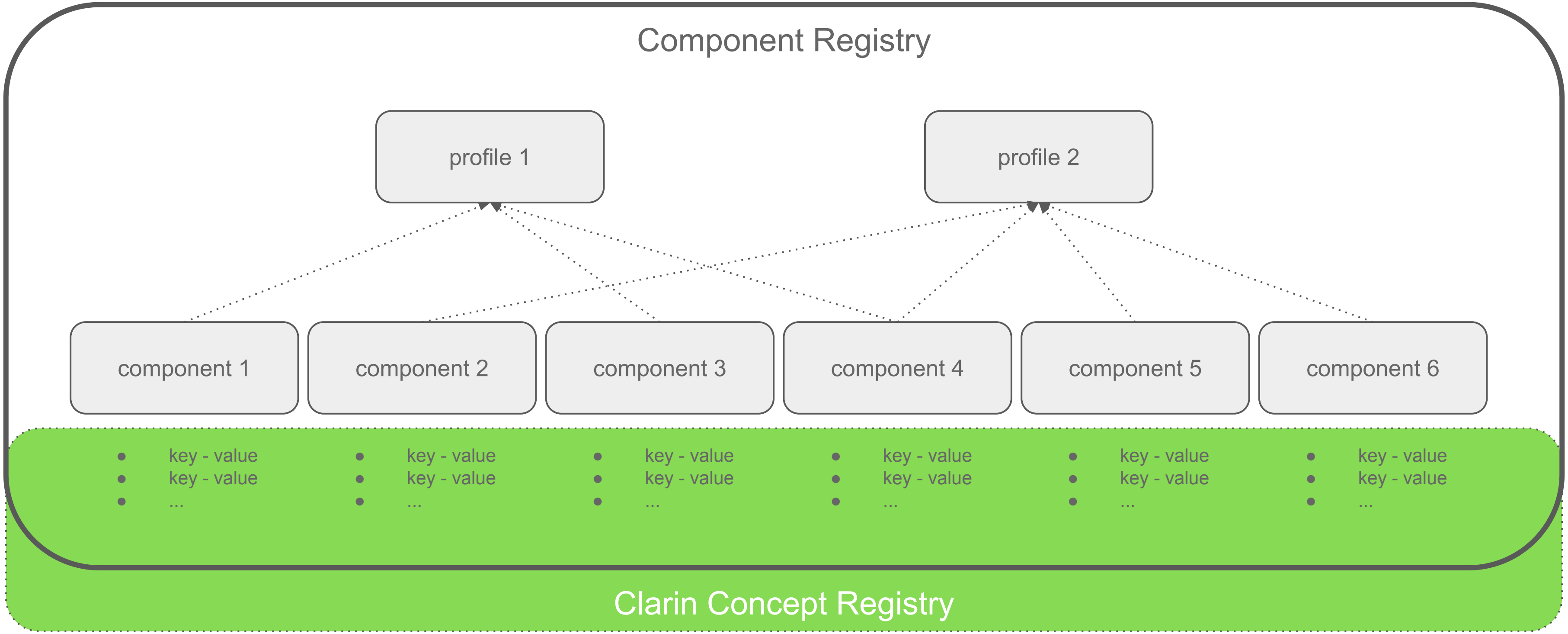
SUMO classes identification ↓

- cognitive schema matching ↓
- TRAJECTOR and LANDMARK identification ↓

Set of semantic schemes, plWordnet, SUMO ontology, Serdel (plWordnet to SUMO mapping, Names to plWordnet and SUMO mapping

CMDI

Component MetaData Infrastructure (CMDI) → framework to describe and reuse metadata blueprints



CMDI profile example

link

CMDI benefits

- architectural freedom
- powerful exploration and search possibilities over a broad range of language resources
  - Virtual Language Observatory
  - Meertens Institute CMDI search engine
- the Component Registry supporting CLARIN Concept Registry (CCR) when creating a Concept Link in the profile/component editor

CMDI benefits

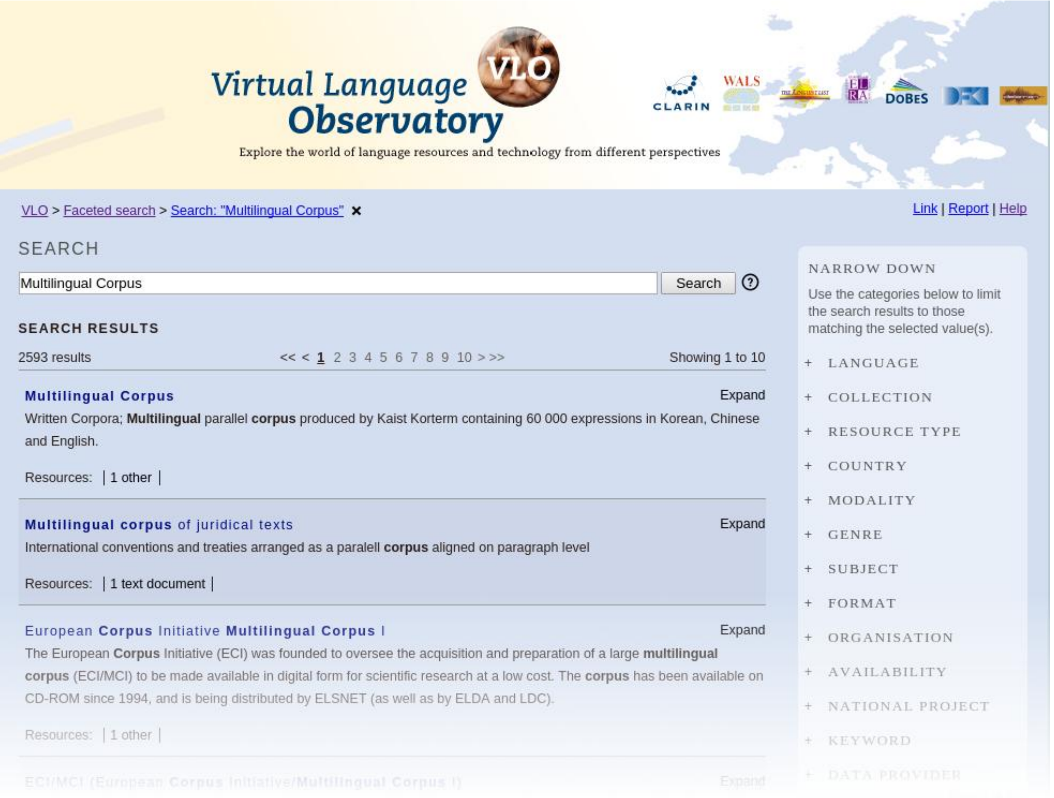
- architectural freedom
- powerful exploration and search possibili e resources
  - Virtual Language Observatory
  - Meertens Institute CMDI search engine
- the Component Registry supporting CLARIN Concept Registry (CCR) when creating a Concept Link in the profile/component editor

location country
mime type
genre
sub genre
tag
language ID
language name
language usage

language availability
life cycle status
modalities
organization
project name
project title
resource class
TEI Header type
domain of use
classification code

time coverage
end range
start range
IPR holder
legal owner
availability
rights
source country

CMDI search possibilities



Virtual Language Observatory
VLO
Explore the world of language resources and technology from different perspectives
CLARIN
WALS
ELRA
DOBES
DFKI

VLO > Faceted search > Search: "Multilingual Corpus" ✖

Link | Report | Help

SEARCH

Multilingual Corpus

SEARCH RESULTS

2593 results

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Showing 1 to 10

Multilingual Corpus Expand

Written Corpora; **Multilingual** parallel **corpus** produced by Kaist Korterm containing 60 000 expressions in Korean, Chinese and English.

Resources: | 1 other |

Multilingual corpus of juridical texts Expand

International conventions and treaties arranged as a paralell **corpus** aligned on paragraph level

Resources: | 1 text document |

European **Corpus** Initiative **Multilingual Corpus** I Expand

The European **Corpus** Initiative (ECI) was founded to oversee the acquisition and preparation of a large **multilingual corpus** (ECI/MCI) to be made available in digital form for scientific research at a low cost. The **corpus** has been available on CD-ROM since 1994, and is being distributed by ELSNET (as well as by ELDA and LDC).

Resources: | 1 other |

ECI/MCI (European **Corpus** Initiative/**Multilingual Corpus** I) Expand

NARROW DOWN

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

- + LANGUAGE
- + COLLECTION
- + RESOURCE TYPE
- + COUNTRY
- + MODALITY
- + GENRE
- + SUBJECT
- + FORMAT
- + ORGANISATION
- + AVAILABILITY
- + NATIONAL PROJECT
- + KEYWORD
- + DATA PROVIDER

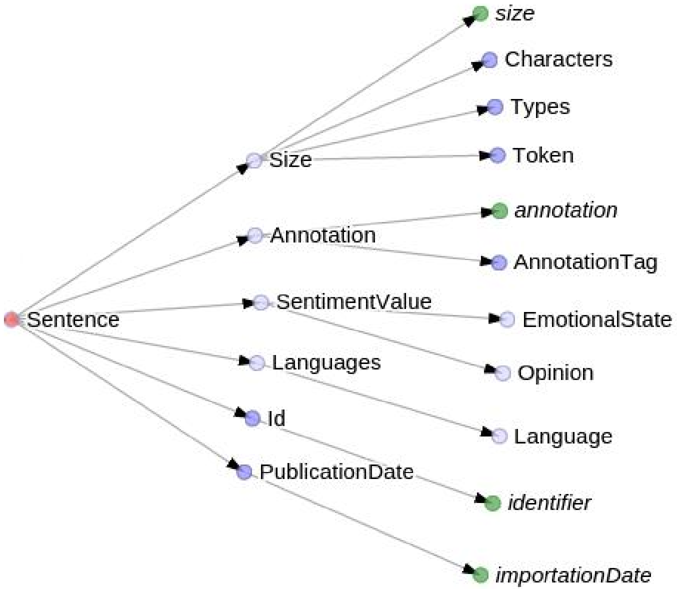
CMDI in annotation process

CMDI instances are assigned to a document/text/recording…

But what about the lower levels?

CMDI in annotation process

Sentence component/profile



CMDI structure:

The <Resources> element

This section of the CMDI **file** enumerates files which are parts of or closely related to the described resource

Thank you for your attention