Towards Mapping Thesauri onto plWordNet

Marek Maziarz and Maciej Piasecki

G4.19 Research Group
Computational Intelligence Department
Wrocław University of Science and Technology

& CLARIN-PL

clarin-pl.eu

Politechnika
Wrocławska

Agenda

- plWordNet
- Lexical resources and terms
- Polish vocabulary in lexical resources
- Proposed mapping
- Checking term lexicality

plWordNet - a wordnet for Polish

- Since 2005
- Constructed manually from scratch
- Broad coverage of Polish vocabulary
  - 219 000 synsets
  - 191 000 lemmas (literals)
  - 285 000 lexical units
  - >100 000 glosses
  - >100 000 usage examples
- Dense lexical net, rich relation system
  - 100 relation types and subtypes
  - 650 000 relation instances
- Mapped onto Princeton WordNet
  - >200 000 relation instances

Goal

- to design a linking mechanism between plWordNet and a rich cloud of heterogeneous terminological and ontological resources, as well as Linked Open Data,
- and next to develop an efficient method for building this mechanism in a semi-automated way.

Mappings from plWordNet

mapping	**relation type**	**instances**
plWN-WN	*l-synonymy*	44K
plWN-WN	*l-near-synonymy*	7K
plWN-WN	*l-hyponymy*	125K
plWN-Wikipedia	*exactMatch*	55K

Table 1: Selected mapping relations from plWN to Princeton WordNet and to *Wikipedia*.

Lexical resources

- There are various types of resources containing *terms* (especially – language vocabulary):
  - **glossaries** are simple, subject oriented lists of terms and their meanings;
  - **dictionaries** expand term lists with sense/concept textual definitions, often beyond one given subject domain;
  - **taxonomies** arrange vocabulary (terms) by hierarchical relations,
  - **thesauri** are based on more complex relation system and lexical description,
  - **lexical databases** - like WordNet - use a couple dozen lexico-semantic relations between (sets of) senses, mixing them with textual definitions and other properties (registers, corpus frequency, valence frames etc.),
  - **formal ontologies** concentrate on concepts, instances (individuals) and logical definitions, but list also *labels*.

Terminology, Terms and Lexical Units

- The word *term* is polysemous.
- Two similar meanings:
  1. 'a piece of specialist terminology'
  2. 'either a piece of language lexicon or a free word-combination'
- Lexical resources deal with $terms_2$ (in the broader sense)

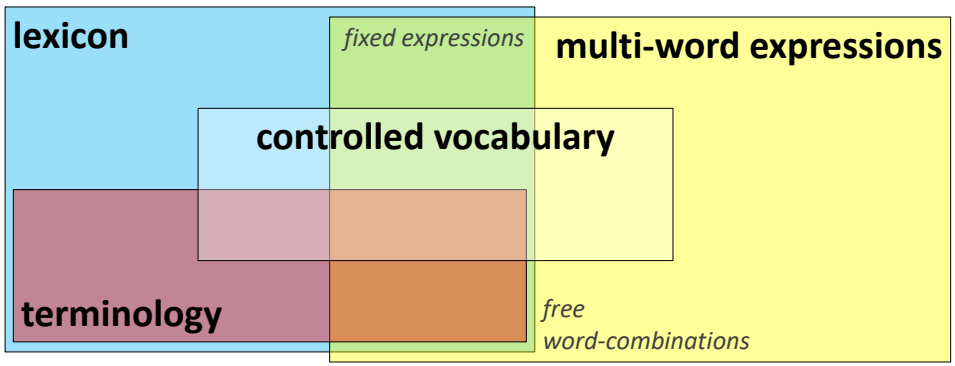


Figure 1: Relations between lexicon, terminology, multi-word expressions and controlled vocabulary. *Controlled vocabulary* is housed often by thesauri and formal ontologies, it uses $terms_2$ in set meanings (polysemy is avoided).

Lexical resources with Polish vocabulary

resource	licence	Polish terms$_2$	external links
DBpedia^s	CC-BY-SA	∼1M	>100K
PNLSH^m	non-commercial	∼100K	20K
IATE^s	sim. to CC-BY	72K	>100K
Agrovoc^s	CC BY-NC-SA	29K	50K
MeSH^{m,s}	sim. to CC-BY	28K	10K
Eurovoc^s	sim. to CC-BY	10K	10K
Gemet^s	sim. to CC-BY	5K	7K
UDC^s	CC-BY-SA	2.5K	0.5K
Sternik^s	sim. to CC-BY	1.7K	—
Digizaurus^s	CC-BY-NC	0.6K	—

Table 2: Lexical resources containing Polish terms$_2$: *s* in superscript marks resources available in SKOS format, *m* represents MARC 21 format.

Important formats
SKOS

SKOS Most interesting resources are published in SKOS

types of information Concepts, schemes (group of concepts), labels, semantic relations, mapping relations

taxonomic relations `broader` and `narrower` link concepts which are hierarchically super-/subordinate or in part/whole relation.

mapping relations `exactMatch` links strict equivalents, `closeMatch` links to a less precise counterpart, `broadMatch`/`narrowMatch` points to the external concept which has broader/narrower extension

Important formats

MARC 21

MARC 21 Native format for the Library of Congress Subject Headings (and other subject heading systems)

field 080 provides counterparts from Universal Decimal Classification

field 082 links to Dewey Decimal Classification

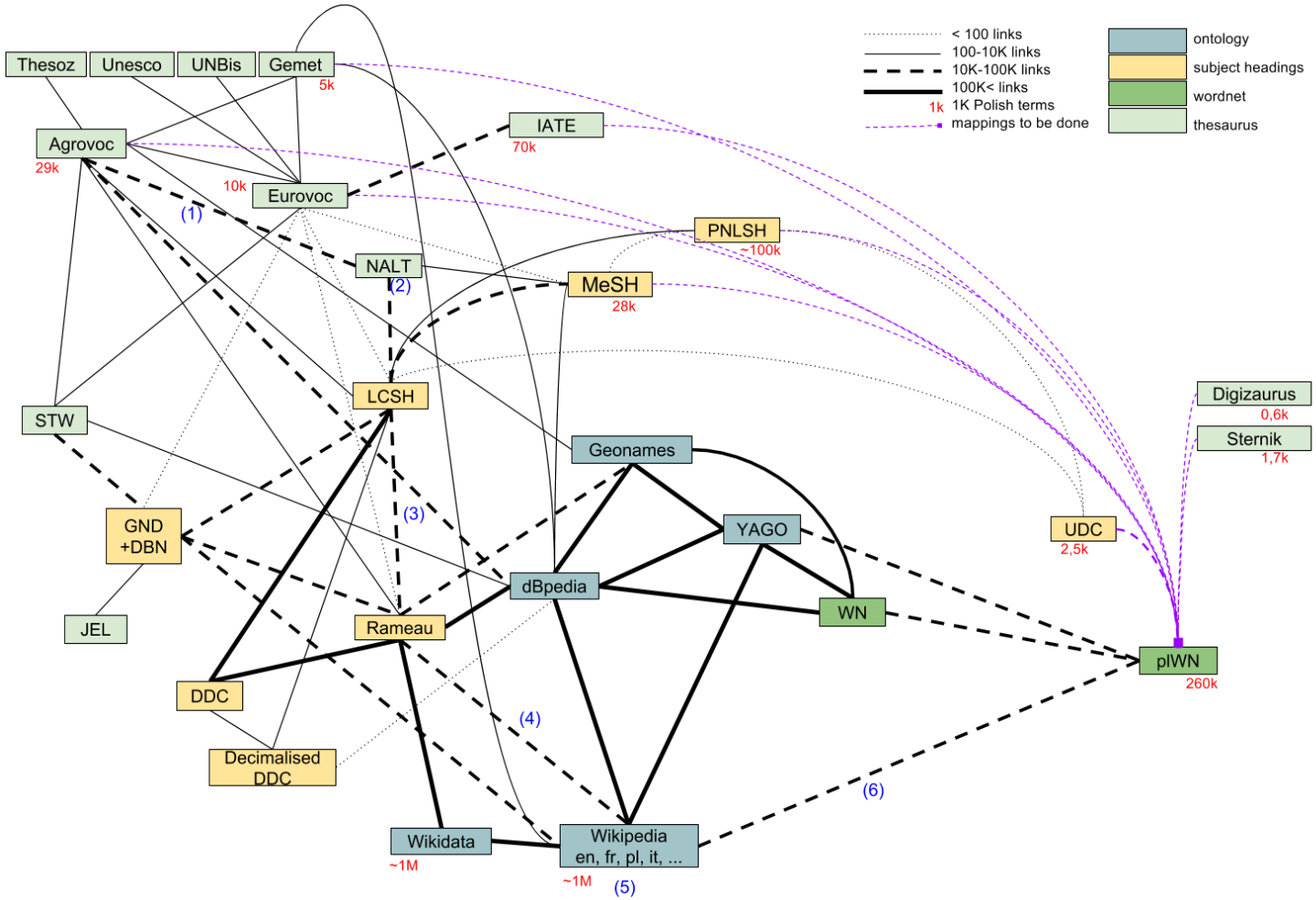
fields 150/450 give preferred and alternative labels (respectively).

field 550 lists all internal semantic relations within a given subject headings system

field 650 gives equivalents in distinct resources: "0" stands for LCSH, "2" – MeSH.

plWordNet & lexical resources

Linking potential – Polish perspective.



The proposed mapping

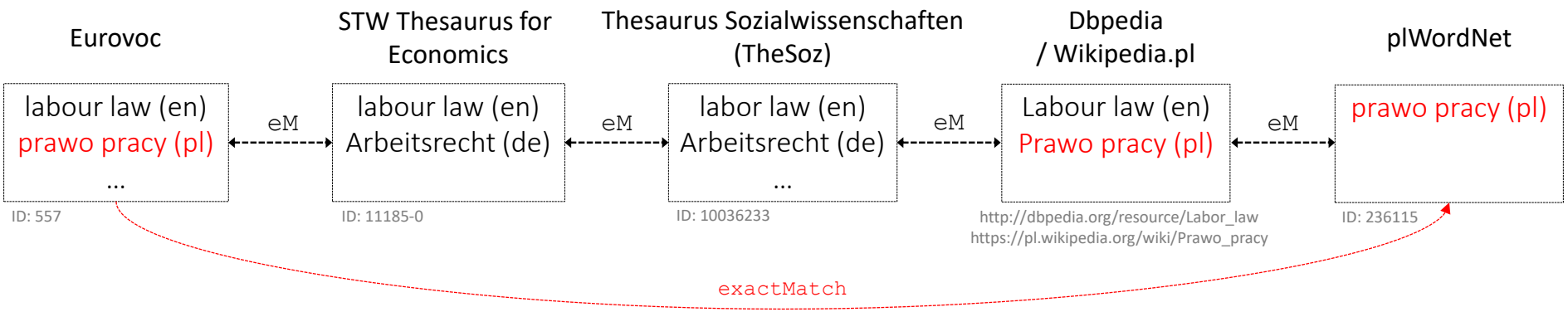
Main objectives

- All these lexical resources are interlinked, composing a pretty complex resource net.
- We want to find a path through it to plWordNet.
- Mapping onto Princeton WordNet and Wikipedia gives plWordNet its own window on the world.
- Thesauri lacking Polish labels may be equipped with Polish equivalents through vocabulary *propagation*.
- The next step will be running an automatic matching algorithm (giving suggestions for linguists).
- Lexicalised $terms_2$ might be introduced into plWordNet.

The proposed mapping

Finding a path

- Let's examine an example path from Eurovoc to plWordNet through Wikipedia.



Linking the concept 'labour law' from Eurovoc with the plWordNet synset $\{\text{prawo pracy}_1\}$; `eM` stands for `exactMatch`.

The proposed mapping

Vocabulary propagation

- Many lexical resources do not contain Polish terms$_2$.
- We may propagate Polish vocabulary using SKOS `exactMatch` relation.
- Consider `exactMatch` chains: Eurovoc $\rightarrow$ STW $\rightarrow$ TheSoz (the two latter do not possess Polish labels).

mapping	**relation type**	**instances**
Eurovoc-STW	*exactMatch*	2262
Eurovoc-STW	*closeMatch*	369
STW-Thesoz	*exactMatch*	3021

Table 3: Mappings from Eurovoc to STW & from STW to TheSoz through direct links.

The proposed mapping

Vocabulary propagation (2)

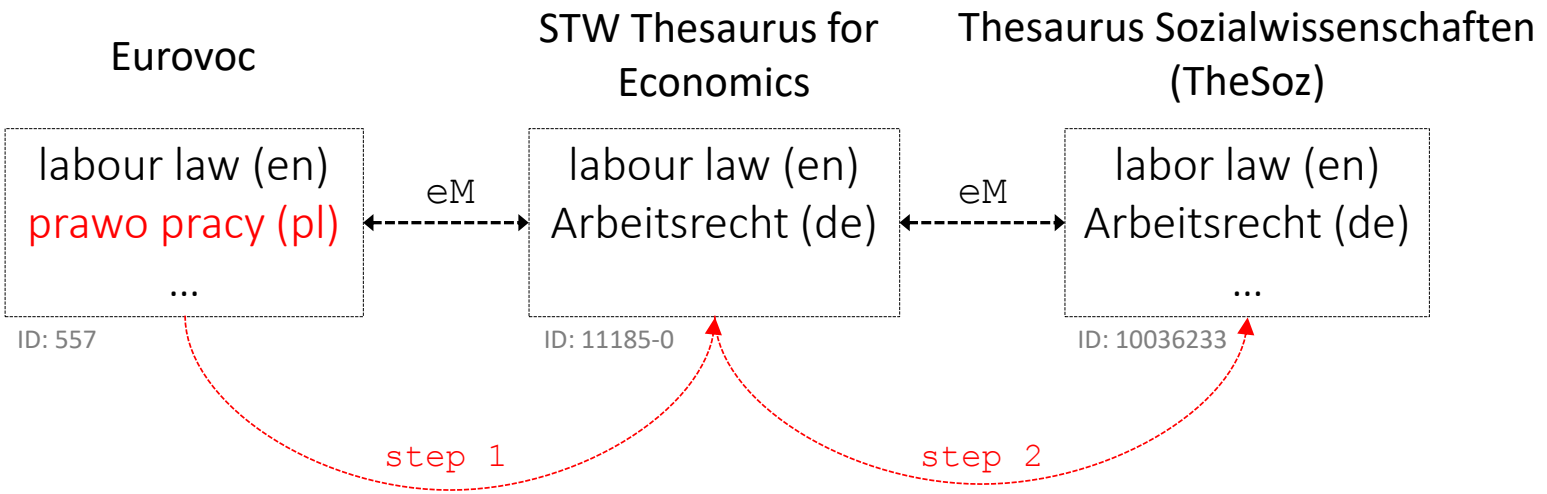


Figure 5: Iterative process of translating labels of the concept 'labour law' in STW and TheSoz; `eM` stands for `exactMatch`.

The proposed mapping

Hybrid approach

- Establishing new `exactMatches` using existing mappings
- Suggesting potential links with an automatic matching algorithm
  - The implementation of relaxation labelling algorithm[1]
  - Used in the large-scale mapping plWordNet onto Princeton WordNet
  - Capable of linking isolated thesauri (like Sternik or Digizaurus)
- Constant evaluation of mapping quality
  - $>$100K Polish terms$_2$ to link
  - checking all potential links too costly
  - manual inspecting sampled examples

[1] Kędzia et al. 2013, cf. Daude et al. 2003

The proposed mapping

Hybrid approach (2)

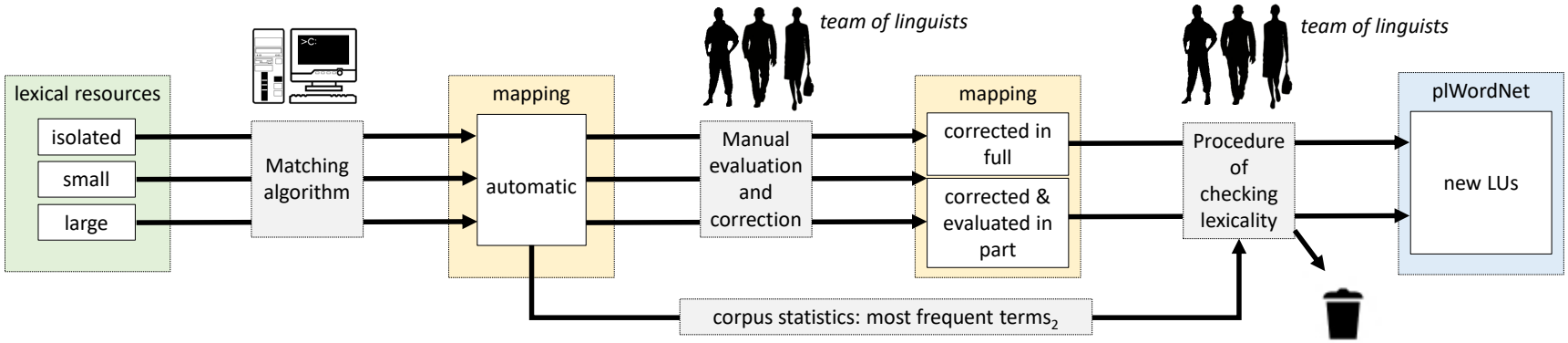


Figure 6: Semi-automatic mapping lexical resources onto plWordNet.

Non-lexicalised terms$_2$

- Many terms$_2$ occurring in lexical resources are not (lexicalised) common nouns.
- They might be semantically transparent, free word-combinations: “Three-wheeled vehicle”, “Two-wheeled vehicle”, “Electric two wheeled vehicle” [DBpedia].
- They may even contain *conjunctions*, *prepositions* or *commas*: cf. “regions and regional policy”, “water management in agriculture” [Eurovoc], “Chemicals and Drugs”, “Influenza, Human” [MeSH].
- They may be given in plural: cf. “Tanks” [DBpedia], “Virus Diseases’ [MeSH], “Organisms” [Agrovoc].
- They may be *proper names*: “Spear of Destiny (video game)”, “Spear Of Destiny Computer Game” [DBpedia].

The procedure of checking $term_2$ lexicality

Introducing $terms_2$ into plWordNet

1. Is X a $term_2$? Y: **next**, N: **end**
2. Can X serve as a noun in a sentence? Y: **next**, N: **end**
3. Is X a proper name? Y: **end**, N: **next**
4. Is X already introduced into plWN? Y: **end**, N: **next**
5. Is X a plurale tantum? Y: **goto** 6, N: **next**
6. Is X a plural form? Y: **end**, N: **next**
7. Is X a MWE? Y: **next**, N: **introduce** X
8. Is a conjunction / comma a part of X? Y: **end**, N: **next**
9. Is X semantically compositional? Y: **next**, N: **introduce** X
10. Does X belong to terminology? Y: **introduce** X, N: **next**
11. Does X exhibit syntactic irregularity? Y: **introduce** X, N: **end**

Legend: **next** means 'go to the next step of the procedure', **goto** denotes jumping to the specific step, **end** = 'X is not a lexical unit', **introduce** = 'add a $term_2$ to plWordNet'.

Perpectives

- Great potential in building a very large network of resources around wordnets
- Expansion of the network utilising the existing high quality manual mapping of plWordNet onto WordNet
- Improvement of a wordnet-based WSD that works better with larger and denser network
- Basis for a method of the automated assignment of descriptive keywords to texts and support for extraction of keywords from texts
- Automated semantic indexing of digital research repositories
- Different applications in Digital Humanities and Social Sciences

Thank you very much for your attention!

http://clarin-pl.eu

http://nlp.pwr.edu.pl

http://plwordnet.pwr.edu.pl