Towards Mapping Thesauri onto plWordNet

Marek Maziarz, Maciej Piasecki
G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Technology, Wrocław, Poland
marek.maziarz|maciej.piasecki@pwr.edu.pl

Abstract

plWordNet, the wordnet of Polish, has become a very comprehensive description of the Polish lexical system. This paper presents a plan of its semi-automated integration with thesauri, terminological databases and ontologies, as a further necessary step in its development. This will improve linking of plWordNet into Linked Open Data, and facilitate applications in, e.g., WSD, keyword extraction or automated metadata generation. We present an overview of resources relevant to Polish and a plan for their linking to plWordNet.

1 Introduction

After more than 12 years of continuous development plWordNet – the wordnet of Polish – with the version 3.0 emo (Maziarz et al., 2016) has become a very comprehensive description of the Polish lexical system including: 197,721 synsets, 179,125 lemmas and 260,214 Lexical Units (henceforth LUs[1]) described by about 650,000 relation links. It provides also a very good coverage of large corpora of Polish, cf (Maziarz et al., 2016). This is much more than it could have been expected at the beginning, especially if we take into account that plWordNet has been constructed from scratch on the basis of the corpus-based wordnet development method (Maziarz et al., 2013). Moreover, plWordNet has been also manually mapped onto Princeton WordNet on the synset level to a very large extent (>200K mapping relation instances) and onto Wikipedia on the LU (sense) level (55K mapping relations). Selected statistics are presented in Tab. 1. It includes also emotive annotation for more than 31,000 LUs (Zaśko-Zielińska et al., 2015).

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

Table 1: Mappings from plWordNet to Princeton WordNet and to *Wikipedia*.

The question is whether it is the final stage of the development of a wordnet of Polish, or more generally, an example of the final stage of a wordnet in general? The immediate answer is no. A complete wordnet is a moving target that evolves along two dimensions: increasing understanding of the effective use of a wordnet as a tool in describing the lexical system of the natural language, and growing expectations of the wordnet applications developers. In this paper we are going to focus on the latter. Wordnets have to compete with statistical models that are relatively easy to extract from very large corpora. However a wordnet is (or must be) a trustworthy language resource of high quality, providing description of the lexical meanings and the lexical system. Its advantage is in description of infrequent lemmas and LUs that is beyond the scope of Distributional Semantics methods (including word embeddings). Next, an appropriate, high quality means of linking a wordnet with knowledge resources must be provided to facilitate its applications in WSD, keyword and semantic meta-data extraction from text, semantic text classification etc. Our goal is to design a linking mechanism between plWordNet and a rich

[1] A lexical unit is defined her technically as a triple: ⟨Part of Speech, lemma, sense id.⟩

cloud of heterogeneous terminological and ontological resources, as well as Linked Open Data (LOD), and next to develop an efficient method for building this mechanism in a semi-automated way. In this paper, we focus on linking with terminological resources as a natural extension to the wordnet.

2 Terminology, Terms and Lexical Units

2.1 Ontologies, thesauri, wordnets

The word *ontology* means many things. Most prominent semantic distinction is between 'metaphysics' vs 'a specific kind of computer science object', however, there is a huge debate on how to define the word in the latter sense:

> "Ontology has become, at least for a time, a prevalent buzzword in computer science. An unfortunate side-effect is that the term has become less meaningful, being used to describe everything from what used to be identified as taxonomies or semantic networks all the way to formal theories in logic." (Pease, 2011).

According to (Roussey et al., 2011) several types of ontologies can be distinguished in relation to their components and structure, including:

Formal ontologies focus mainly on instances (individuals), concepts and their logical definitions (a.k.a. *axioms*) combine logic operators and quantifiers with relations between concepts, and thus enable reasoning.

Software implementation driven ontologies "provide conceptual schemata whose main focus is normally on data storage and data manipulation, and are used for software development activities, with the goal of guaranteeing data consistency" (*ibidem*).

Linguistic ontologies[2] focus mainly on labels and relations between them:

- *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 (hypo-/hypernymy, type-/instance, broader/narrower, see (Mitkov and Matsumoto, 2004)),
- *thesauri* are based on a more complex relation system: apart from sub-/superordinate relations also other lexico-semantic links are involved, cf (Currás, 2010),
- *lexical databases* - like WordNet - use a couple dozen lexico-semantic relations between (sets of) senses (concepts), mixing them with textual definitions and other properties (register labels, frequency information, semantic domains, valence frames etc.).

Information ontologies – used by humans in project development processes – aim at capturing relations between concept instances in diagrams in order to clarify the ideas of collaborators.

We adopt here the term *formal ontology* in the meaning: "a formal, explicit specification of a shared conceptualization" (Studer et al., 1998).[3] The term *lexical resource* will be used instead of *ontology* (in its broader sense) for all types of computer science objects comprising concepts, their instances, properties, labels and relations between them in various configurations.

Several phenomena arise in vocabulary formalisation. Mapping between concepts and their lexicalisations is not one to one. Existence of near-synonymy and sense vagueness cause that there is no clear cut between many semantically related word senses, and they often overlap. Only subtle differences constitute the distinctions (Fellbaum, 2011). This is captured by a concept of *near-synonymy*, a relation that links word senses close in meaning, being equivalents (interchangeable) in some, but not in all contexts.

In fact, also mapping from words to concepts is not straightforward due to polysemy. Especially many frequent words possess two or more meanings, which is an unusual situation in a formal ontology.

[2] *Lexical* ontologies lack formalization which is characteristic property of formal ontologies, but the former might be comparable to the latter in taxonomic parts (like biology vocabulary), cf. (Hirst, 2009).

[3] The word "conceptualization" means here 'an abstract, simplified view of the world that we wish to represent for some purpose' (Guarino et al., 2009). This knowledge ought to be shared by a group of people / a community (e.g., specialists in a given field), and the specification should be so intuitive that most stakeholders could agree with it.(Vrandečić, 2009). Moreover, an ontology should be formally specified and formal logic (usually first order logic or Description Logic) should be used for description purposes to avoid any ambiguities (Prévot et al., 2010).

Structural (lexical) gaps are also problematic: the mental lexicon does not lexicalise all concepts people have in mind, so there appear gaps in lexical taxonomies (Vossen, 2004).

Natural language in not a *formal* language and the formalization of a vocabulary, even the formalization of relational dictionary, is not an easy task. Consider group / mass nouns *armament – weaponry* and try to ascribe them a relation type. Would it be meronymy or hyponymy?

Lexicon is *not* a formal ontology, nevertheless

> "a formal ontology without natural language labels attached to classes or properties is almost useless, because without this kind of grounding it is very difficult, if not impossible, for humans to map an ontology to their own conceptualization, i.e. the ontology lacks human-interpretability."(Völker et al., 2007), after (Hirst, 2009)

2.2 Terms and lexical units

Dictionaries, thesauri, wordnets and formal ontologies in a way deal with vocabulary. A formal ontology uses words as labels that help people to find out the meanings of ontology concepts. A dictionary concentrates on words – describes words, their meaning, grammatical properties and usage. Thesauri and wordnets interlink words and their senses into a lexical net, encoding their description by lexico-semantic relations.

Apart from words, all these resources tend to house some multi-word expressions (MWEs), either fixed (lexicalised) or free. The distinction between what is a part of a vocabulary (what is a multi-word LU) and what is a free syntactic word combination (a collocation)[4], although not entirely clear, is valid for dictionaries, terminological thesauri, and some wordnets (plWordNet, Germanet). However, in formal ontologies, many domain thesauri and WordNet, words, fixed and free phrases are mixed up. For instance, in the thesaurus of European Union *Eurovoc* we may find free word combinations: *regions and regional policy* or *water management in agriculture*. Similarly in *MeSH* we spot MWEs *Chemicals and Drugs* and *Virus Diseases* (plural). In WordNet we notice word combinations *wheeled vehicle* and *horse-drawn vehicle*[5]. Many entries occurring in these lexical resources are domain specific. This leads us to the problem of demarcation between terminology and ordinary phrases and words. The distinction lies in the specialist nature of terminology and the natural provenance of ordinary vocabulary. Terminology is known mostly to specialists, while ordinary language is spoken by all of us.[6]

In ISO 1087-1 *term* is a "verbal designation of a general concept in a specific subject field". (Wright and Budin, 2001, p. 325) defines *terminology* as "the (structured) set of concepts and their representations in a specific subject field". These two exemplar definitions suggest that concepts dominate over their lexical manifestations within terminology. Conceptual structure of a theory may enforce morphological shape of words (like in chemistry nomenclature) or can influence formation of MWEs (e.g. in biological taxonomy).

Despite the dissimilar provenance of ordinary and specialist vocabulary, they do not differ with regard to their relation to meaning:

> "[T]he relationship between concept and terms is *formally equivalent* to the relationship between meaning and words." (...) "The traditional theory of terminology [claims] that the concept is the meaning of the term". (Kageura, 2002)

Terms consist of phonemes, they have their morphemes, inflect like ordinary words or are composed of words like ordinary compositions and have inflection like ordinary phrases. Like ordinary lexemes they do have their meanings. Since they "are [formally] indistinguishable from words" (Sager 1998/99, after: (Kageura, 2002)), we treat terminology as a part of the lexicon.

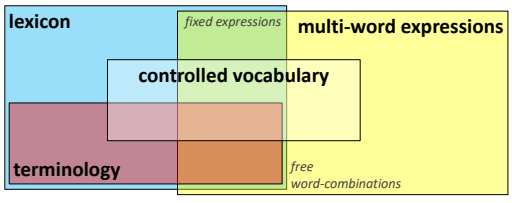

Figure 1: Relations between lexicon, terminology, multi-word expressions and controlled vocabulary.

[4] We call semantically or syntactically fixed MWEs *multi-word lexical units* (MWLUs, cf. (Zgusta, 1967)). According to some linguists semantic or syntactic fixedness of MWEs is merely a symptom of being a part of one's mental lexicon, see (Svensén, 2009; Müller, 2015; Sprenger, 2003).

[5] In Germanet such MWEs are called 'artificial'.

[6] These are specialists that invent new scientific terms, their discussion how to define terms is the important part of scientific activity. On the contrary, ordinary language has no father and evolves spontaneously.

In Fig. 1 we present the relationships between *lexicon*(blue rectangle), *terminology* (red) and *word-combinations* (yellow). By the white one we mark the *controlled vocabulary*.

The controlled vocabulary could be found in thesauri (like *Eurovoc*), ontologies (like *SUMO*) and in subject headings systems (like *Library of Congress Subject Headings*, *LCSH*, or *MeSH*). It consists of specialist terms, ordinary words, multi-word LUs and free word-combinations, sometimes it uses plural forms representing a given category. An important feature of a controlled language is its avoidance of semantic ambiguities:

> "Word or phrase indexing and symbolic surrogation systems require some sort of controlled vocabulary – an artificially constructed language in which the ambiguities of natural language are reduced or, ideally, eliminated. A controlled vocabulary is an organized list "of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation or search." Controlled vocabularies have two primary objectives: (1) to represent concepts systematically and (2) to facilitate comprehensive searching of a body of information." (Wallace, 2007)

It is worth to emphasise that *term* is used not only in the meaning 'a unit piece of terminology', but also in a broader sense. It may denote every single label/lemma (word or MWE), being an entry of an ontology, a thesaurus, a wordnet or any other lexical resource. All kinds of language expressions from Fig. 2.2 could be described by this word. In this paper, if we use *term* in its broader sense, we will write it down with the plus mark in a superscript (so, *term*$^+$), and if we want to refer to the narrower sense ('terminology unit'), we will write it without a plus (*term*).

plWordNet has concentrated on the Polish *lexicon*, avoiding free word combinations and proper names. Our definition of multi-word LUs points to the phenomena of lexicalisation and terminologisation (Maziarz et al., 2015).

3 Lexical resources vs. plWordNet

Polish vocabulary outside plWordNet could be found in many electronic lexical resources. We describe them below in three groups: (1) subject headings systems, (2) controlled vocabulary thesauri (of the EU, UN and US), and (3) *Wikipedia*.

3.1 Subject headings

There are five available subject heading systems comprising Polish terms$^+$, and the biggest one is the Polish National Library Subject Headings.

Polish National Library Subject Headings (PNLSH) is a descriptor system based on the model of Library of Congress Subject Headings. It has reached circa 100K subject terms$^+$ and still grows. PNLSH makes use of MARC 21 format, like LCSH.

MeSH, Medical Subject Headings, is the US National Library of Medicine's controlled vocabulary for medicine. Polish translation was prepared by Main Physicians' Library in Warsaw, Poland. It gives 28K Polish terms$^+$. MeSH is mapped onto LCSH, Snomed or US National Agricultural Library Thesaurus.

Universal Decimal Classification (UDC) core was published on CC-BY-SA licence and translated into Polish by Polish National Library. The UDC core itself is linked to LCSH and through it to Dewey Decimal Classification (DDC) and MeSH.

Sternik is yet another subject headings system designed by Polish National Library. Housing terminology of bibliography and cataloguing, it gives also translations to English. It is equipped with the associative relation *related term*, definitions and alternative labels. Unfortunately, *Sternik* is isolated and has no links to external resources.

Digizaurus is a small thesaurus carefully designed by Polish Digitalization Inter-Museum Group *DigiMuz* for museum collection description in the field of *material*. It comprises 0.6K terms$^+$ organised into taxonomy (obtainable in SKOS). Digizaurus is also an isolated resource, like Sternik.

resource	licence	terms$^+$	links
PNLSH^m	NC	∼100K	20K
MeSH^{m,s}	NC	28K	10K
UDC^s	CC-BY-SA	2.5K	0.5K
Sternik	sim. to CC-BY	1.7K	—
Digizaurus^s	CC-BY-NC	0.6K	—

Table 2: Subject headings systems for Polish. The label "terms$^+$" denotes Polish labels in each vocabulary, "links" describes an approximate number of mapping instances to external resources (for all terms$^+$, including Polish), "NC" means 'non-commercial', the letter *s* in superscript marks resources available in SKOS RDF format, *m* represents MARC 21 format.

3.2 Thesauri

IATE, InterActive Terminology for Europe, is a large thesaurus developed collectively by the community of translators and institutions of the EU. It comprises 8.6 million terms$^+$ in 24 languages. Polish vocabulary numbers 72K terms$^+$.
Eurovoc is an open licence thesaurus describing activities of the EU. It provides terminology in 26 languages, also in Polish (10K terms$^+$). Eurovoc has mappings to multiple other thesauri (given in SKOS), inter alia: Agrovoc, Gemet, LCSH, STW Thesaurus for Economics or UNESCO Thesaurus.
Agrovoc was created by Food and Agriculture Organization (FAO) of the United Nations. It is pretty well linked to many external resources, among them to Eurovoc, Gemet, Rameau, STW, Geonames, Thesos and 16 open datasets related to agriculture. Polish translation was done by Central Agricultural Library and comprises 29K terms$^+$.
Gemet, GEneral Multilingual Environmental Thesaurus, was developed by European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA). It contains multilingual environment terminology (5K Polish terms$^+$) and is a reference thesaurus in this field.

resource	licence	terms$^+$	links
IATE^s	sim. to CC-BY	72K	>100K
Agrovoc^s	CC BY-NC-SA	29K	50K
Eurovoc^s	sim. to CC-BY	10K	10K
Gemet^s	sim. to CC-BY	5K	7K

Table 3: Polish controlled vocabularies in thesauri.

3.3 Wikipedia

Wikipedia.pl and their byproducts – YAGO or dBpedia — comprise hundreds of thousands of Polish terms$^+$. The whole vocabulary is structured with Wikipedia category system. YAGO expanded this system merging it with WordNet. *Wikipedia* is developed by the community of volunteers.

resource	licence	terms$^+$	links
Wikipedia	CC-BY-SA	$\sim$1M	>100K

Table 4: *Wikipedia* comprises most Polish terms$^+$.

4 Linking Potential

All these lexical resources are interlinked, composing a quite complex resource net. We want to find a path through it in order to establish mappings between them and plWordNet. We will exercise two main formats: SKOS and MARC 21.

4.1 Formats and alignment

Most resources described in this paper are recorded in SKOS RDF and in MARC 21 (for subject headings). Other relevant formats e.g., of WordNet, of Wikipedia, of dBpedia and of YAGO, will not be discussed, due to space limit.

SKOS RDF. Simple Knowledge Organization System[7] provides "specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web." and uses the Resource Description Framework (RDF). In SKOS RDF we have following types of information:

- Concepts: "units of thought – ideas, meanings, or (categories of) objects and events".
- Concept groups - *schemes* (thesauri or microthesauri grouping concepts) and *collections* (smaller groups of concepts).
- Labels: expressions used in a natural language to refer to concepts. One label is *preferred*, all the others are *alternative* forms.
- Notes: describes concepts in various ways, for instance, *definitions* are verbal descriptions of term$^+$'s meaning.
- Semantic relations: describe concepts in the net of semantically closest concepts. Relations *broader* and *narrower* link concepts which are hierarchically super-/subordinate or in part/whole relation.
- Mapping links between a parent thesaurus and external resources are encoded with *.*Match* relations: *exactMatch* links strict equivalents, *closeMatch* links to a less precise counterpart in one external resource, *broadMatch*/*narrowMatch* points to the external concept which has broader/narrower extension, *relatedMatch* denotes other semantic relations – they are crucial in our task.

MARC 21. MARC (MAchine-Readable Cataloging) 21 is a data format (ISO 2709) used for cataloguing and bibliographic description. It is used

[7] https://www.w3.org/2004/02/skos

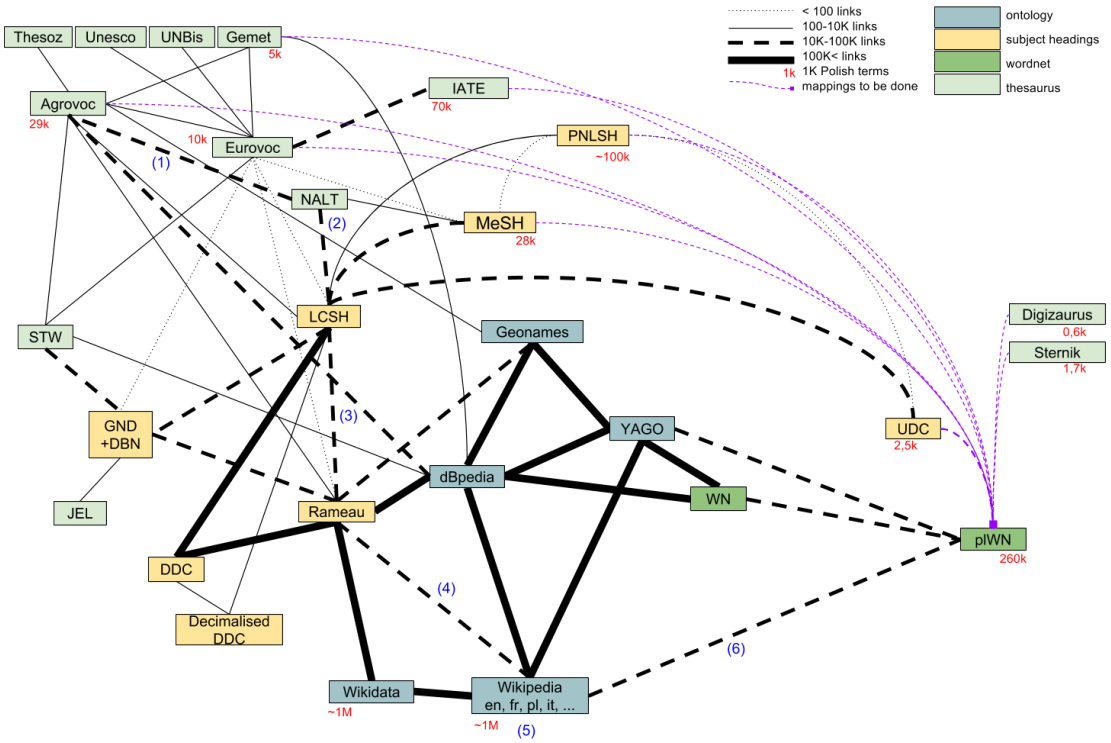

Figure 2: Linking potential of the existing lexical resources – Polish perspective.

by the Library of Congress in its famous subject headings that makes it popular. MARC provides various *fields* of which the most important for us are:

- Field `080` provides counterparts from UDC, while `082` links to DDC.
- Fields `150` and `450` gives 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.

4.2 Vocabulary 'propagation'

Existing mappings between lexical resources give an opportunity not only to align Polish vocabulary between two separate thesauri, but also to provide translations for not-translated terms^{+}. Thesauri lacking Polish labels may be equipped with Polish equivalents. Let us call it vocabulary *propagation*.

We plan to propagate the vocabulary iteratively. At first, we will use direct links between resources to label equivalents with Polish labels. Then we are going to use such translated lexical resources to translate resources that are linked to them. Thus Polish vocabulary would spread across the net of lexical resources. In each step we will proceed only with translations of direct equivalents.

Direct equivalents. Let us look at existing Eurovoc - STW Thesaurus for Economics and Eurovoc - Gemet mappings (see Tab. 5 and Fig. 2). In Eurovoc SKOS RDF we find 2262 *skos:exactMath* links to STW and half as many to Gemet. Some of them have Polish labels in Eurovoc. STW does not, and Gemet does. Consider the Polish label *prawo pracy* 'labour law' in *Eurovoc*, its concept (ID: 557) has the exact match in STW (labelled *labour law*) and the exact match in *Gemet* (labelled with *prawo pracy*).

mapping	relation type	instances
Eurovoc-STW	*exactMatch*	2262
Eurovoc-STW	*closeMatch*	369
Eurovoc-Gemet	*exactMatch*	1294

Table 5: Mappings from Eurovoc to STW & Gemet through direct links.

In step 1 we give Polish labels to all concepts that have an exact or close match in a mapping from any labelled with Polish terms$^+$ thesaurus.

Indirect equivalents. To exemplify how we plan to establish indirect links let us discuss the case of a Polish label for 'blood protein disorders' in Agrovoc (ID: c 969): *Zaburzenia białek krwi* (preferred label[8]). Since we may link the label to the National Agricultural Library Thesaurus (NALT) concept 'blood protein disorders' (ID: 18150), we may also take advantage of NALT-LCSH mapping existence (cf. Tab. 6). The concept has the exact equivalent in LCSH *Blood protein disorders* (ID: sh 85015013).

mapping	relation type	instances
Agrovoc-NALT	*exactMatch*	26520
NALT-LCSH	*exactMatch*	8501
NALT-LCSH	*closeMatch*	2755

Table 6: Mappings from Agrovoc to US National Agricultural Library Thesaurus (NALT) & from NALT to LCSH through direct links.

Even longer paths. We may go with the Agrovoc even beyond LCSH. In Fig. 2 one may find a possible way from Agrovoc to plWordNet (marked with blue numbers): Agrovoc $-1\rightarrow$ NALT $-2\rightarrow$ LCSH $-3\rightarrow$ Rameau $-4\rightarrow$ Wikipedia francophone $-5\rightarrow$ Polish Wikipedia $-6\rightarrow$ plWordNet. Let us trace the whole path with the concept 'blood pressure' from Agrovoc (ID c 967).

(1) The concept has the Polish label *Ciśnienie krwi* (`prefLabel`; the alternative label *Obniżone ciśnienie*, lit. 'low blood pressure', is not considered here). It points to NALT 'blood pressure' (ID: 18146) via *exactMatch*. **(2)** NALT 'blood pressure' then is matched with LCSH 'Blood pressure' (ID: sh 85015010), again with the *exactMatch* relation. **(3)** From LCSH we jump right to French National Library subject headings *Rameau* and 'Pression artérielle' (ID: cb11976295t). The *closeMatch* was used here.[9] **(4)** Now we go with *exactMatch* to French Wikipedia to the article *Pression artérielle*[10] and then **(5)** to Polish Wikipedia article *Ciśnienie tętnicze* (='artery pressure'[11].) **(6)** Since plWordNet is widely linked to Polish Wikipedia with *exactMatch*, we may finally establish link from Agrovoc ID: c 967 *Ciśnienie krwi*, *blood pressure* to the plWordNet synset {ciśnienie tętnicze 1}.

The above example raises the question on the quality of such long chains. The longer the path is, the more probable the relation is distorted. Is *ciśnienie krwi* 'blood pressure' a real synonym of *ciśnienie tętnicze* 'arterial pressure'? Fortunately, we do not have only one way to choose from a given resource to plWordNet. Thanks to the mapping between plWordNet and Princeton WordNet our path bifurcates. We may choose a route from the WordNet through ontologies YAGO and dBpedia to Rameau. This gives us rare occasion to verify different links and check their consistency.

4.3 Hybrid approach

When the iterated process of vocabulary propagation is done, we will have some Polish terms$^+$ introduced into different lexical resources, as well as, many matching relation instances. Of course, links to plWordNet synsets are of special importance and the whole process will focus on them.

Prompt algorithm. The next step will be running an algorithm giving suggestions to linguists. It takes into account the already established links as constraints. We plan to utilize the implementation of relaxation labelling algorithm (used successfully in plWordNet-WordNet mapping (Kędzia et al., 2013)). The algorithm can handle also linking isolated resources (like Sternik or Digizaurus).

Assessing quality of the mapping. The automatic algorithm will suggest potential links. We may expect more than 100K new terms$^+$, so assessing quality of the automatic mapping will be a challenge. Mappings from small resources (e.g. Gemet) could be checked fully by plWordNet editors, and manual checking of the mappings of isolated thesauri (Digizaurus and Sternik) is a must. However, automatic matching from larger resources, like Polish Wikipedia or PNLSH, will be too big for a complete manual verification. The proposed process is presented in Fig. 3.

After checking and correcting automatically generated links, linguists will also check lexical-

[8] Please, note that – according to SKOS guidelines – only preferred labels are linked by the exactMatch relation.

[9] Please note that: (a) the blood pressure is usually measured in arteries, (b) *closeMatch* is supposed to serve well only on short distances (one link, see SKOS definition).

[10] `https://fr.wikipedia.org/wiki/Pression_artérielle`

[11] `https://pl.wikipedia.org/wiki/Ciśnienie_tętnicze`

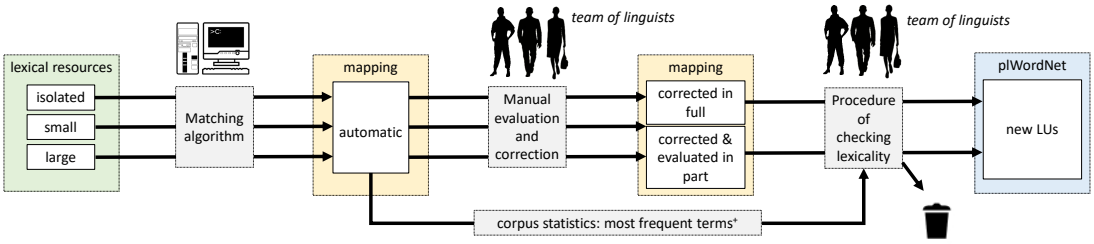

Figure 3: Semi-automatic mapping lexical resources onto plWordNet. The matching relation verification will be done in full (for small and isolated thesauri) or in part (for large resources). Linguists may check also lexicality of all verified in the preceding phase terms$^+$ plus some of high corpus frequencies.

ity of terms$^+$ taken from isolated or small lexical resources, and a sample of terms$^+$ from large resources together with the most frequent ones. We estimate that verification of 1K automatic links and assessing their lexicality will take altogether one person-month, e.g. preparing the mapping of Sternik, would take two person-months, while Agrovoc circa 30 person-months. In order to remain consistent with most of our thesauri (Agrovoc, Digizaurus, Eurovoc, Gemet, IATE, MeSH and UDC) relation types from the SKOS format will be utilized. Linguists will choose semantically closest counterparts from plWordNet, whether they will be exact or close equivalents (*exactMatch*, *closeMatch*), or synsets which have broader or narrower meaning (*broadMatch*, *narrowMatch*).

Listing 1: Introducing terms$^+$ into plWN

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

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$^+$ to plWordNet', **term**$^+$ denotes either a word or a MWE being a part of a lexical resource.

Introducing LUs into plWordNet. The mapping will give us a unique opportunity to expand plWordNet with new LUs. This will be done in two phases. Firstly, we will check it at the same time as the matching relation accuracy evaluation. Secondly, we will test those terms$^+$ that are frequent in a reference corpus.

As we have shown in Sec. 2.2, many terms$^+$ occurring in lexical resources are not lexicalised. Among them there are entries containing *conjunctions*, *commas*, being *free* word-combinations and *proper names*, or given in *plural*. We propose the following algorithm designed for plWordNet editors (Listing 1) to asses a given term$^+$ as a LU.

The 10 filtering rules help sifting through non-lexicalised language expressions. At the end, lexicalised terms$^+$ are introduced into plWordNet.

5 Perspectives

The presented overview and mapping method show a great potential in building a very large network of resources around plWordNet. The network can be even more expanded with LOD utilising the existing high quality manual mapping of plWordNet onto WordNet. The primary application will be improvement of a wordnet-based WSD that works better with larger and denser network. Next, it will be a basis for a method of the automated assignment of descriptive keywords to texts and will support extraction of keywords from texts. Both methods will be first used in automated semantic indexing of digital research repositories, and next in different applications in Digital Humanities and Social Sciences. For such applications possibility of finding associations between texts and specialist terms is crucial and can be done via the created complex network.

Acknowledgment Works funded by the Polish

Ministry of Science and Higher Education within CLARIN-PL Research Infrastructure.

References

[Currás2010] Emilia Currás. 2010. *Ontologies, Taxonomies and Thesauri in Systems Science and Systematics*. Chundos Publishing, Oxford.

[Fellbaum2011] Christiane Fellbaum. 2011. Wordnet (and why it's not an ontology). In Adam Pease, editor, *Ontology: A Practical Guide*, pages 71–73. Articulate Software Press, Angwin, CA, USA.

[Guarino et al.2009] Nicola Guarino, Daniel Oberle, and Steffen Staab. 2009. What is an ontology? In Steffen Staab and Ruder Studer, editors, *Handbook on Ontologies*. Springer, second edition.

[Hirst2009] Graeme Hirst. 2009. Ontology and the lexicon. In *Handbook on ontologies*, pages 269–292. Springer.

[Kageura2002] K. Kageura. 2002. *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. Terminology and lexicography research and practice. J. Benjamins Pub.

[Kędzia et al.2013] Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka, and Konrad Przybycień. 2013. Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies*. to appear.

[Maziarz et al.2013] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In G. Angelova, K. Bontcheva, and R. Mitkov, editors, *Proc. Int.l Conf. on Recent Advances in Natural Language Processing*, Hissar, Bulgaria.

[Maziarz et al.2015] Marek Maziarz, Stan Szpakowicz, and Maciej Piasecki. 2015. A procedural definition of multi-word lexical units. In *RANLP*, pages 427–435.

[Maziarz et al.2016] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plwordnet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL.

[Mitkov and Matsumoto2004] Ruslan Mitkov and Yuji Matsumoto, 2004. *Handbook Of Computational Linguistics*, chapter Lexical Knowledge Acquisition. Oxford University Press.

[Müller2015] Peter O. Müller, 2015. *Multi-word expressions*. De Gruyter Mouton.

[Pease2011] Adam Pease. 2011. *Ontology - A Practical Guide*. Articulate Software Press, Angwin, CA.

[Prévot et al.2010] Laurent Prévot, Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, and Alessandro Oltramari. 2010. Ontology and the lexicon: a multi-disciplinary perspective. In *Ontology and the Lexicon. A Natural Language Processing Perspective*. Cambridge University Press.

[Roussey et al.2011] Catherine Roussey, Francois Pinet, Myoung Ah Kang, and Oscar Corcho, 2011. *An Introduction to Ontologies and Ontology Engineering*, pages 9–38. Springer London, London.

[Sprenger2003] Simone Sprenger. 2003. *Fixed expressions and the production of idioms*. Ph.D. thesis, Max Planck Instituut voor Psycholinguïstiek.

[Studer et al.1998] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. 1998. Knowledge engineering: Principles and methods.

[Svensén2009] Bo Svensén. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press.

[Völker et al.2007] Johanna Völker, Pascal Hitzler, and Philipp Cimiano, 2007. *Acquisition of OWL DL Axioms from Lexical Resources*, pages 670–685. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Vossen2004] Piek Vossen, 2004. *Handbook Of Computational Linguistics*, chapter Ontologies. Oxford University Press.

[Vrandečić2009] Denny Vrandečić, 2009. *Ontology Evaluation*, pages 293–313. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Wallace2007] Danny P. Wallace. 2007. *Knowledge Management: Historical and Cross-disciplinary Themes*. Libraries Unlimited.

[Wright and Budin2001] S. E. Wright and G. Budin. 2001. Handbook of terminology management: Application-oriented terminology management. John Benjamins, Amsterdam and Philadelphia.

[Zaśko-Zielińska et al.2015] Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for Polish. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing – RANLP'2015*, pages 721—-730, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

[Zgusta1967] Ladislav Zgusta. 1967. Multiword lexical units. *Word*, 23(1-3):578–587.